PREDICTORS OF SUCCESS IN AIR FORCE INSTITUTE OF TECHNOLOGY RESIDENT MASTER'S DEGREE PROGRAMS: A VALIDITY STUDY

James R. Van Scotter, Captain, USAF

LSSR 4-83

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY (ATC)
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

PREDICTORS OF SUCCESS IN
AIR FORCE INSTITUTE OF TECHNOLOGY
RESIDENT MASTER'S DEGREE PROGRAMS:
A VALIDITY STUDY

James R. Van Scotter, Captain, USAF

LSSR 4-83

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deleterious information are contained therein. Furthermore, the views expressed in the document are those of the author(s) and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the Air Training Command, the United States Air Force, or the Department of Defense.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER LSSR 4-83	2. GOVT ACCESSION NO. AD-A134391	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PREDICTORS OF SUCCESS IN AIR FORCE INSTITUTE OF TECHNOLOGY RESIDENT MASTER'S DEGREE PROGRAMS: A VALIDITY STUDY		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James R. Van Scotter, Captain, USAF		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Systems and Logistics Air Force Institute of Technology WPAFB, OH		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Department of Communication AFIT/LSH, WPAFB, OH 45433		12. REPORT DATE September 1983
		13. NUMBER OF PAGES 118
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Approved for public release: IAW AFR 190-17.

LYNN E. WOLAVER

Dean for Research and Professional Development

Air Force Institute of Technology (ATC)

Wright-Patterson AFB OH 45433

15 SEP 1983

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Selection
Officer personnel
Psychological tests
Schools
Psychological measurement

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Thesis Chairman: Guy Shane, PhD

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

The effectiveness of a selection process is directly related to the strength of the relationships between eligibility criteria and measures of performance. Criterion-related validity research provides a means to establish the nature and strength of predictor/criterion relationships. Predictors used by the Air Force Institute of Technology (AFIT) in selecting military officers and civilian employees for graduate school include admission test scores and undergraduate grade-point-averages. Using a sample containing 2170 cases which spanned the period from 1977 to 1982, predictor/criterion relationships were demonstrated for AFIT's in-residence master's degree programs. The differences in key predictor/criterion relationships between 17 graduate programs were statistically tested. As a result, it was found that some master's degree programs could be statistically combined with others to enhance prediction. Prediction models were developed using multiple regression. These models were shown to be superior to the present selection process. Correlation matrices, program groups, and prediction models are presented.

LSSR 4-83

PREDICTORS OF SUCCESS IN AIR FORCE INSTITUTE OF TECHNOLOGY RESIDENT MASTER'S DEGREE PROGRAMS: A VALIDITY STUDY

A Thesis

Presented to the Faculty of the School of Systems and Logistics

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Logistics Management

By

James R. Van Scotter, BS
Captain, USAF

September 1983

This thesis, written by

Captain James R. Van Scotter

has been accepted by the undersigned on behalf of the faculty of the School of Systems and Logistics in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN LOGISTICS MANAGEMENT

DATE: 28 September 1983

Guy S. Shane

COMMITTEE CHAIRMAN

John A. Muller

READER

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

The United States Air Force has been a leader in the design, development, a management of new technologies throughout its history. The evolution of military air power has been paralleled by a growth in the size and complexity of the Air Force. By consistently exploiting the military applications of technological advances, the Air Force has been able to counter the threats posed by potential adversaries and increase its own capability to support the national objectives of the United States. In meeting this challenge, the Air Force has made a strong committment to advanced technical and management education. The Air Force Institute of Technology (AFIT) is a concrete example of that committment. Through AFIT's programs, military officers and civilian employees of the Department of Defense are sponsored in both undergraduate and graduate level programs in management, engineering, and other related disciplines.

Although AFIT administers a wide range of educational programs, its in-residence master's degree programs are of particular interest.

> These programs are designed to give selected officers the ability to analyze and solve complex technical and managerial problems faced by the Air Force and the Department of Defense. (United States Air Force Manual 50-5, Volume I, para 4-9 (a), 1981)

This ur que emphasis provides students with many of the skills necessary for successful performance at the higher levels of the Air Force organization and opens career opportunities for them. Thus, both the individual students and the Air Force benefit.

Class size limitations, and the need to maximize the return on its investment require the Air Force to employ a selective admissions policy. Only those officers whose academic abilities, motivation, and job performance indicate a high probability of success are admitted to AFIT's master's degree programs. To differentiate between those students who are likely to succeed and those who are not, AFIT has established basic eligibility criteria. Minimum requirements vary somewhat from program to program but, in general, AFIT requires an undergraduate grade-point-average (UGPA) of 2.5 or better (on a 4.0 scale) and Graduate Record Examinations (GRE) Verbal, and Quantitative scores totalling 1000, or a Graduate Management Admissions Test (GMAT) score of 500 or better (Air Force Institute of Technology, 1982).

Undergraduate grade-point-averages have been widely used as eligibility criteria in graduate and professional

schools. However, in recent years, grade inflation, a wide range of grading practices, and the advent of non-traditional degree programs has made this index increasingly difficult to interpret (Knapp and Hamilton, 1978). As a result, the use of other indicators, and especially the use of standardized tests such as the GRE and GMAT, has become more and more important.

Professionally prepared standardized tests can provide valuable information about the skills and aptitudes of potential graduate students. Information about the distribution of test scores for recent examinees is made available to graduate schools by test publishers. This enables graduate school admissions departments to evaluate students from a wide variety of backgrounds on a common scale, and to compare the performance of each individual with national averages.

When test scores are used to differentiate between applicants, an underlying assumption is that they measure attributes that are strongly related to academic performance. Another assumption is that scores occur in a continuum and can be ranked. Following this logic, it is further assumed that high scores indicate high potential, average scores indicate average potential, and low scores low potential.

These assumptions may or may not be valid. The degree to which they are valid in predicting success in a specific situation is a critical concern. Many factors can influence the accuracy of such predictions, but they can be reduced to the test itself, individual differences, and differences in how "success" is defined.

> A test cannot be valid in general; it is valid for a purpose. Indeed, a test may be both valid and invalid. For example, skill in algebra may be a valid predictor of science and math grades, but it may not be valid for history or english. (Green, 1981)

Green's point is well taken. A test must measure pertinent skills if it is to be useful. The relationship between a predictor and a criterion (measure of success) is expressed in terms of the predictor's criterion-related validity. Test users must have evidence of the criterion-related validity of the test to insure that their decisions are based on relevant information.

There is also another issue. Those involved with the use of tests for selection must be aware of ethical considerations. In an effort to establish some ethical guidelines for test users and publishers, the American Psychological Association published a handbook entitled: <u>Standards for Educational and Psychological Tests and Manuals</u> (1966). Its purpose is to provide a common framework for evaluating tests and test materials. The reason for its development follows:

Almost any test can be useful for some functions and in some situations, but even the best test can have damaging consequences if used inappropriately. Therefore, primary responsibility for the improvement of testing rests on the shoulders of test users.(American Psychological Association, 1966, p.6)

This general admonition was followed with a more specific discussion of the criterion-related validity issue.

Local collection of evidence on criterion-related validity is frequently more useful than published data . . . In cases where criteria differ from one locality to another or from one institution to another, no published data can serve all localities. For example, the validity of a certain test for predicting grades at a college with a unique kind of curiculum may be quite different from the published validity of the same test that was based on more conventional colleges. (American Psychological Association, 1966, p.18)

Of course it can be argued that until the criterion-related validity of a test has been demonstrated for a specific purpose, a user cannot ethically rely on published data. This type of argument, along with practical concerns about the effectiveness of the GRE and GMAT as predictors, has led many graduate schools to sponsor local validity studies. The Educational Testing Service (ETS), developer and publisher of the GMAT and GRE has been even more specific in its recommendations.

It is incumbent on any institution using GMAT scores in the admissions process that it demonstrate empirically the relationship between test scores and measures of performance in its academic program. (Graduate Management Admissions Council, 1982)

> All parties to the development of the Graduate Record Examinations (GRE) Program, from the outset, have recognized the need for empirical evidence regarding the predictive validity of the GRE tests and other preadmissions variables. (Wilson, 1979)

Independent researchers have reached similar conclusions:

> Each professional school should carry on continual research on the effectiveness of its selection procedures and various other aspects of its total program. Selection procedures need to be empirically validated, since one cannot assume that they will be effective in one situation if they have been so in others. (Furst, 1950)

Practical consideration are very important. The value of any selection instrument is directly related to the savings it provides the organization. According to AFIT's 096CR financial report for 1981, the average costs for sponsoring graduate students in the Engineering and Logistics Management Schools were $82,892.68 and $67,258.66 respectively (Air Force Institute of Technology, 1981). There are many ways to view this investment, but no matter what your perspective is, it is reasonable to assume that the Air Force gains more from graduates than non-graduates. If you view non-graduation as a total loss, it is evident that even a small improvement in the selection procedure could result in significant savings.

> When tests are used in the selection process, it should be on the basis of demonstrated improvement in the selection with the test score over selection without the test score. (Womer, 1968, p.57)

Local validity studies can furnish this information and may also point out better ways of combining the various predictors. A tailor-made prediction model can be developed for a particular situation.

Admissions eligibility criteria can become outmoded. If this happens the efficiency or the selection process can be seriously affected.

> A test with significant criterion-related validity five or ten years ago may not have the same relationship today. This will be particularily true whenever there is any change in the criterion. A college that becomes more selective over a period of years may change its grading practices enough to alter the predictive validity of a college aptitude test. (Womer, 1968, p.61)

<u>Problem statement</u>

GRE and GMAT test scores are used in determining the eligibility of candidates for AFIT's resident master's degree programs in the School of Systems and Logistics and in the School of Engineering. While other factors are also considered, GRE and GMAT test scores are heavily weighted by the AFIT Registrar's staff.

Simply stated, the problem is that the relationships between the various indicators of student potential and academic success in these programs have not been demonstrated. More specifically, the validity of the GRE and GMAT as predictors of academic performance for most

of these programs has not been established, nor has the validity of existing selection procedures been analyzed. Empirical research is clearly called for. Until this research is accomplished, no basis exists for criticizing or endorsing AFIT evaluation procedures. All we can say conclusively is that we do not know whether or not the evaluation process is accurate. An empirical study may provide support for the AFIT Registar's selection process, including its use of GRE and GMAT scores, or it may suggest that other methods could be more useful. In either case, it should furnish a basis for evaluating past, present, and future admissions practices.

The primary purpose of this study is to evaluate the criterion-related validity of the GRE and GMAT and other variables as predictors of success in AFIT resident master's degree programs. To establish a basis for comparison, the validity of the present selection process was investigated. Finally, prediction models were developed and their effectiveness compared with the historical accuracy of AFIT admissions decisions.

Background

The criterion-related validity of the Graduate Record Examinations (GRE) and Graduate Management Admissions Test (GMAT) in predicting student success in

graduate schools has been the subject of many studies. The GRE has become firmly established as a device for evaluating the relative academic potential of prospective graduate students throughout a wide range of academic disciplines. The GMAT, which is designed for use by business and management schools, has also become an important tool in graduate student selection (Hecht and Powers, 1982). Both tests have known reliability, and are general enough to measure the knowledge, aptitudes, and skills of a wide variety of individuals from different educational backgrounds (Educational Testing Service, 1981).

The GRE and GMAT are standardized tests with norm and scale scores. Standardization refers to the administration, apparatus, and scoring methods associated with the use of the measurement device. Educational Testing Service (ETS) insures through carefully controlled formal administration procedures, that each time a test is given the same specific steps are followed by the test proctor. Each version of a test is identical in appearance to other versions of the same test. Each has the same number of questions, the same type of answer sheets, and each follows the same format. Each version is analyzed to insure its content parallels that of other versions (Educational Testing Service, 1981).

The term "scaled score" as it is used by ETS, refers to the practice of using a reference group to establish a scale against which the performance of subsequent examinees can be measured. According to ETS, the reference group for the GRE consisted of a large group of college seniors from eleven undergraduate institutions who took both the GRE verbal and quantitative subtests in 1952. The mean score for this entire group was set to equal 500, and a standard deviation was set at 100. Through statistical manipulation of test score data, ETS sets the means and standard deviations of subsequent groups of examinees to the same parameters. ETS asserts that comparisons between the scores of two (or more) examinees is useful and valid when consideration is given to errors of measurement. That is, ETS is careful to point out that small differences in test scores are relatively meaningless (Educational Testing Service, 1981).

Tests, like other tools, are designed with specific purposes in mind. As an aptitude test, the GRE is designed to measure the effects of learning that occurred over a relatively long period of time under relatively uncontrolled conditions. Its purpose is to predict performance. This can be contrasted with the use of achievement tests to measure the learning and skills that a person has acquired in a more structured formal setting

(Anastasi, 1976, Educational Testing Service, 1981). The GMAT, which measures knowledge in a specific area to a greater extent than the GRE does, is more of an achievement test than an aptitude test. It is important to realize that there is no absolute distinction between the two types of tests, since similar items appear in both. The distinction is based on the use of the scores from the tests rather than any inherent differences in the tests themselves.

Both the GRE and the GMAT are divided into subtests. The GRE has three subtests; verbal, quantitative, and analytical. The analytical subtest was added to the GRE in 1977, and scores for it have been reported since 1978. ETS cautions that this test should not be used for decision making until its validity can be demonstrated. The analytical test is designed to measure an individual's ability to reason in a logical way, to reach sensible conclusions, and to identify the important factors in a situation. The purposes of the verbal and quantitative subtests are to measure aptitudes in those areas. The GMAT contains subtests in only the verbal and quantitative areas.

On the surface, quantitative measures are easier to interpret than subjective measures. They fit more easily into decision criteria formulae. Certainly, it is easy to

pick the higher of two scores. It is much more difficult to make a decision based on individual traits such as motivation, persistence, and maturity, though most people would agree that these factors contribute to an individual's performance. In fact, subjective appraisals have been shown to be less effective than decisions based on statistical measures in many circumstances because of variability among raters and differences in criterion definitions (Sawyer, 1966). In addition, quantitative measures tend to lend credibility to a selection process (Furst, 1950, Marston, 1971). In terms of practical results, the use of test scores has, in general, improved the efficiency of many organizations both inside and outside the educational arena (Travers, 1956).

A substantial body of research deals with the effectiveness (or ineffectiveness) of the GRE as a predictor of success in graduate education. Similar research of the GMAT is limited by comparison. Criticisms of the predictive validity of the GRE and the GMAT have centered around the low correlations that have been found in studies of the relationships between these predictors and the criterion of academic success as measured by graduate grade-point average (GGPA). The key factors contributing to the low correlations are explored below.

Reliability

The concept of test reliability deals with the accuracy of the measurement.

> In its broadest sense, test reliability indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. (Anastasi, 1976)

The Educational Testing Service is responsible for developing and managing the GRE and GMAT programs. ETS has consistently demonstrated that the reliability of both tests is above .90 (Hecht and Powers, 1982, Educational Testing Service, 1981). As pointed out by Cureton, reliability is a necessary prerequisite for meaningful validity (Cureton, 1950). For the purpose of predicting student performance in graduate school, a high measure of reliability increases our confidence that a given prediction is meaningful.

Validity

Validity is concerned with what tests measure. In general, it can be described as the usefulness of the measurement. Criterion-related validity, the central concept in prediction, is a combination of two of these; concurrent validity and predictive validity. Criterion-related validity emphasizes the relationship (i.e., correlation) between a test score (predictor) and

some other measure of behavior, the criterion of success (Womer, 1968). The criterion of success is some future performance that is of interest. When considering prediction of academic performance, the criteria of graduate grade-point average and graduation/non-graduation have frequently been employed. These criteria, and others, are used to improve the accuracy with which schools select students that are likely to succeed (Womer, 1968).

Brogden demonstrated that the correlation coefficient represents the proportional improvement in selection that results from the use of a predictor over what would be expected in a selection based on the criterion alone. He interpreted this "as showing that the correlation coefficient is a direct index of predictive efficiency" (1946). Brogden argues convincingly that decision makers should consider the improvement in selection obtained through the use of predictors in light of the costs associated with obtaining and interpreting them. Brogden's point is that the users of a test are responsible for validating its utility in both economic and predictive terms. Womer makes the same argument. In discussing the use of standardized tests in selection, he states: "The development of local validity studies is the best possible approach to criterion-related validity." (1968, p.61)

If any positive correlation between the predictor and the criterion is achieved, predictions based on the predictor will be more accurate than random chance. Although decision makers would prefer perfect prediction, validity coefficients are usually less than .60 in practice (1968, p.61). Traxler noted that:

> In view of the restricted range of talent usually represented in correlations between test scores and marks at the graduate level, correlations in the neighborhood of .50 may be regarded as satisfactory. (1952, p.476)

Validity coefficients are largest in groups that encompass a wide range of ability levels. In groups where the range of ability is narrrow, validity coefficients tend to be low. As the range of abilities in a group becomes narrower and narrower, it become progressively more difficult to differentiate between members of the group (Chronbach, 1970).

This phenomenon is known as restriction in range. The selection process itself contributes to the problem. As products of successively more and more stringent screeenings, graduate students form a group that is very homogenous compared to the population as a whole. The difficulty in achieving success at higher and higher levels increases, compensating somewhat for the effects of restriction in range. Even so, the usual pattern is for correlation coefficients to decrease as groups become

smaller and more homogenous. In discussing the effects of restriction in range on prediction studies, Furst and Roelfs stated:

> Much of the so-called inconclusiveness has come from low correlations and these, in turn, from persisting conditions, especially restrictions in range owing to selective admissions and attrition. (1979, p.147)

Another factor that tends to reduce validity coefficients is compensatory admissions practices. When students are admitted to graduate school despite low test scores, it is usually because the school is aware of other factors that compensate for the test scores. For instance, students with low test scores may be admitted on the basis of strong UGPA's. If this happens often enough, correlations between GRE scores and GGPA for the body of students are likely to be lower than they would have been had selection been based on GRE scores alone. To the extent that compensatory admissions practices are used, validity coefficents will be reduced (Livingston and Turner, 1982). Robertson and Nielsen (1961) noticed the same effect in their study.

All of this is not meant to suggest that admissions officers should use GRE or GMAT scores exclusively as selection criteria. Other factors may indeed be very useful, they are commonly used in combination with test scores. According to Chronbach (1970), the addition of

other relevant factors to a prediction model will generally improve validity coefficients.

Analytical methods

Prediction is either statistical or clinical. Statistical prediction uses data on past performance of groups to predict future performance. Clinical prediction is judgemental, and may be based on theoretical considerations (Sawyer, 1966). According to Thorndike, the clinical method's only advantage is that it:

> permits combination of scores in other than a linear manner. It permits a maximum of flexibilty in that any pattern, no matter how complex or unique, may be recognized and weighted. For this extreme flexibility to be an advantage, it is necessary (1) that special patterns and combinations of tests, not well represented by a linear combination of scores, be important for success on the job and (2) that there be clinicians available who have the insight to discover those special patterns and the skill to recognize then whenever they appear. We may well be skeptical on both counts, but especially on the second. It represents a severe demand on a clinician's insight to expect him to discover better ways of using test scores than will be given by the best linear combination of those scores, and then to be consistent in identifying and interpreting those patterns when they reappear. (1949, p.201)

Statistical prediction techniques involving multiple regression and/or the use of correlation coefficients have been used in all but one study reviewed in this report. By collecting data on several predictor variables and using stepwise regression or factor analysis,

researchers obtain information about the relative contribution of various predictors to the model. Often, those measures identified as the strongest contributors were entered into the model with no weights applied to them. This practice is known as unit weighting.

Jensen pointed out that:

> Given a set of predictive measures from which it is desired to predict graduate scholastic achievement of different groups, equal powers of prediction should not be arbitrarily given to each or any combination of these variables. Empirical tests should first be made to ascertain differences in group performance based on the predictive and criterion variables and data weight derived for each member of the predictive team (1953, p.328).

By the term "predictive team", Jensen is referring to the group of predictors the researcher considers to be logically related to performance in the criterion. His argument against arbitrary weighting of predictors is sensible, especially when techniques such as multiple linear regression, which assigns weights statistically, are available.

Two studies, Madaus and Walsh (1965), and Covert and Chansky (1977), set out to determine optimal prediction models by dividing large groups into smaller, more specific ones. Through this aporoach they sought to demonstrate that the performance of different groups of people can be best predicted using different prediction models. Their hypothesis was that alternate weighting strategies, based

on the characteristics of subgroups would be more effective than simple unit weighting. Their efforts were successful. Even though the need for research in optimal predictor weighting strategies had been called for by Jensen (1953), efforts to do so have been limited (Covert and Chansky, 1977).

Data collection is statistical (or mechanical) if rules can be prescribed to insure that clinical judgement is not involved. Self-reported and clerically obtained data including psychometric tests, biographical data, and grade reports are examples of statistical data. Interviews and judge's ratings, unless strictly limited to recording pre-specified characteristics are clinical in nature (Sawyer, 1966).

Once data of either type are collected, they are combined in prediction models through step-wise regression or other statistical techniques. These methods identify the predictor with the highest correlation with the criterion and build the prediction model around it. In a step-by-step process the relative contribution of each predictor to the model is evaluated, and the predictors are added to the model in order of their contribution. Once the model cannot be improved through the addition of another predictor, the process is complete. Sawyer noted that when data are collected by both the statistical and

the clinical methods, the advantages of statistical combination is the greatest. In addition, he found that: "The present analysis finds the mechanical mode of combination always equal or superior to the clinical mode..." (1966).

Predictors

The literature on validity studies makes it clear that there are nearly as many approaches to prediction as there are researchers. Although this section focuses on the role of predictors, some discussion of criterion measures and research methodology is inevitably included. Depending on the goals of a particular study, a wide range of predictors have been employed.

Thacker and Williams (1974) reviewed twelve studies of GRE predictive validity which spanned the period from 1957 to 1970. In ten of the twelve studies GGPA was the primary criterion variable. One study (Robertson and Nielsen, 1961) used faculty ratings, and another (Law, 1960) used pass/fail doctoral comprehensives as the criterion. Six of the ten studies using GGPA as the criterion found correlations that were either not statistically significant, or were too low to be used effectively in prediction.

Based on these results, Thacker and Williams'

conclusion that the criterion of GGPA is of doubtful predictive value is not surprising. Other researchers, including Marston (1971) and Nagi (1975) have come to similar conclusions. Thacker and Williams reported that the limited range and inherent variability of the GGPA criterion were partly responsible for this finding. They also noted that: "the use of other measurement criteria has not consistently yielded improved correlations" (Thacker and Williams, 1974). Given the relatively small sample sizes (N was less than 50 in three of the five studies), and the likelihood that other factors also influenced the size of the correlations, this conclusion is not surprising.

Using faculty ratings as the criterion of success, Robertson and Nielsen (1961) arrived at the same conclusions. In their study, nine faculty members each rated fifty students according to their perceptions of the students' ability to complete a psychology doctoral program. The ratings were then combined to form a composite score for each ratee. The mean GRE score correlated with this criterion .27 at the .05 level of significance. The authors concluded that the results were too weak to be used in prediction. However, the combination of mean GRE scores and UGPA in math/science courses correlated .44 with the criterion at the same

significance level, indicating that the combination of the two was a better predictor than was the GRE alone. While this supports Chronbach's observation that increasing the number of predictors will generally increase correlations, it is important to note that Chronbach was referring to a general outcome of adding more variables (information) to a regression model, not a specific situation (Robertson and Nielsen, 1961, Chronbach, 1977).

Nagi used the GRE as a predictor of a dichotomous criterion: completion/non-completion of a doctoral education program. He was unable to find any significant correlations, despite his use of a sample which included thirty non-graduates as well as thirty-three graduates (1975). You might expect that the inclusion of the non-graduates would increase the the heterogeneity of the sample and therefore the size of the correlations achieved. However, this strategy was ineffective. Since the non-graduates were selected for the program, their scores on the predictors were similar to the scores of those who completed the doctoral program. The study may also be criticized for its small sample size, which may have prevented it from achieving more conclusive results.

Camp and Clawson examined the predictive validity of the GRE with respect to the criterion of GGPA. They obtained a correlation of .24 for the total GRE score (the

sum of the GRE verbal and quantitative subtest scores), and a correlation of .27 at the .01 significance level for the verbal subtest alone. They concluded that the results were not strong enough to be useful in predicting success for a group such as the 135 Master of Arts in Counseling candidates they studied (Camp and Clawson, 1979). However, in view of Brogden's (1946) finding that even small improvements in selection can result in significant benefits to the organization, it appears that Camp and Clawson's conclusion was premature. Many studies can be criticised on the same grounds.

<u>Selection based on cut-off scores</u>

To determine the effectiveness of the GRE in discriminating between successful and unsuccessful students, Borg (1963) used a dichotomous criterion. Students were judged successful if their GGPA was equal to or greater than 3.0, and unsuccessful if their GGPA was less than 3.0. To test the hypothesis that successful students can be differentiated from unsuccessful students, Borg created intervals equal to one-half of one standard deviation and computed the number of scores that fell in each of the six intervals he created. He determined that a GRE verbal test cut-off score, established at one-half standard deviation below the mean score for his sample of

175 would have eliminated 72% of the unsuccessful students. The same cut-off score would have had the undesired effect of eliminating 27% of those students that were successful. As a result, admission would have been denied to 41 successful students and 21 unsuccessful students. Based on these findings, Borg concluded that the use of a GRE verbal test cut-off score should not be used at Utah State University (1963).

After analyzing the results of his own predictive validity study and Educational Testing Service reports of other validity studies, Marston (1971) warned against the use of fixed cut-off scores. Marston attempted to predict the publication rates for sixty-four clinical and forty-seven non-clinical psychologists based on the GRE scores they had earned prior to acceptance in graduate school. There was a difference in the correlations he found for the two groups (clinical $r = .01$, non-clinical $r = .27$, $p > .05$). However, the practical value of this information is unclear. Marston's study can be criticized for its small sample size as well as its unreliable and possibly unrealistic criterion.

When several predictors are relevant, a common practice is establish cut-off scores on each of them. The result of using multiple cut-off scores is to eliminate individuals from consideration if their score on any of the

criteria is low. On the other hand, a multivariate linear regression model allows high scores on one predictor to compensate for low scores on another. Compensation can be desirable in situations where a strength in one area can make up for a weakness in another. Multiple cut-off scores are more appropriate in situations where a specific trait, or prerequisite cannot be compensated for by other abilities (Chronbach, 1970, pp.437-438).

One problem with the use of multiple cut-off scores is that there is no analytical method for establishing the minimum acceptable scores. Determining the effect of a single cut-off score is relatively easy, but with multiple cut-off scores the process is neccesarily one of trial and error. The combined effect of multiple cut-off scores creates a non-linear selection model.

> The one case in which we would expect those who were selected by the multiple cutoff procedure to surpass in criterion performance those selected by multiple regression is that in which the relationship of one or more of the tests to the criterion is sharply non-linear. If there is some unique critical score on a particular test below which all or most applicants do poorly on the job and above which a smaller proportion do poorly on the job no matter what their other qualifications, then a procedure which determines that point and establishes a fixed cutoff at that point undoubtedly has advantages. However, in so far as a continuous and approximately linear relationship exists between score on each of the tests and the criterion score of success on the job, no basis exists for choosing a uniquely desirable cutting score. (Thorndike, 1949, p. 198)

In addition, Thorndike points out that multiple cutoff scores provide no information about the degree of suitability of an applicant. In an environment where the intent is to select the best qualified, this method is not particularily useful (1949, p. 199).

Other approaches

In a study of the GMAT, Breaugh and Mann (1981) used discriminant analysis to determine whether or not graduates of an MBA program could be statistically differentiated from non-graduates. The sample consisted of 507 graduate students. Of this group, 266 graduated, 193 voluntarily withdrew, and 48 were terminated for academic defficiency. The authors grouped all non-graduates together.

Breaugh and Mann were able to differentiate the two groups. Student age and GMAT Quantitative subtest scores were the most heavily weighted variables. Their method was 69% accurate in predicting graduation. This was contrasted with the 52% accuracy of the admissions committee. The criteria used by the admissions committee were not mentioned, but statistics on the students were reported. For this sample, the mean UGPA was 2.98, the mean GMAT Verbal score was 31.8 (71st percentile) and the mean GMAT Quantitative test score was 31.7 (also 71st percentile),

indicating that the admissions committee had set "fairly high standards" for applicants (Breaugh and Mann, 1981).

Although the main focus here is on the GRE and GMAT as predictors of success, these variables are seldom used alone. The relationships between a number of other variables and GGPA have been investigated. Of these variables a review of the literature shows that undergraduate grade-point average is the most common. In an analysis of 189 validity studies conducted by ETS between 1975 and 1981, Livingston and Turner found that :

> The combination of GRE scores and undergraduate grades predicts first year grades much more effectively than either the GRE scores alone or undergraduate grade-point average alone. (1982)

Several studies have tested background variables as predictors. Baird (1975) used questionaires to obtain background, self-assessment, and GGPA information on over 2,000 graduate students. He determined that a student's confidence in his abilities and his family background were related to success in business and law schools.

Mehrabian (1969) investigated the effectiveness of a number of variables in predicting success for 266 applicants for admission to a graduate psychology program, and 79 students already enrolled in that program. Prediction criteria of sex, increase in GPA for the last two undergraduate years over the first two years, the rating of the student's undergraduate department, and

research experience were excluded from the final model. Factor analysis showed that those variables did not relate significantly to the success criteria. The success criteria employed were an average evaluation of research competence, average grades in first year statistics courses, and average grades in first year content courses.

Despite the fact that he found little evidence in the literature to support the use of letters of recommendation as predictors, in his own study, Mehrabian reported that they were the second strongest predictor of graduate school succcess. The best single predictor Mehrabian found was the sum of the student's GRE and Miller's Analogy Test (MAT) scores. Although he determined that the UGPA over the last two years of undergraduate school had a stronger relationship to graduate performance than overall UGPA, this predictor was not strong enough to be included in his final model (Mehrabian, 1969).

Mehrabian's criteria are questionable, and his findings have not been replicated. In fact, Lin and Humphreys' study disagrees with Mehrabian's on one point. After analyzing patterns in the undergraduate and graduate school records of over 2,000 students, Lin and Humphreys stated: "There is no evidence that senior grades predict grades in graduate departments more accurately than freshman grades." (1977, p.256)

Lewis examined the relationship between six predictor variables and two criterion measures in the MBA program at the University of Iowa. The predictors included: the number of undergraduate semester hours in business related courses, GPA in those courses, cumulative UGPA, undergraduate major, and Graduate Study in Business Test (GSBT) scores. (The GSBT was a forerunner of the GMAT.) Using stepwise regression, Lewis found that GPA's in business courses and scores on the quantitative portion of the GSBT were the best predictors of GGPA in required MBA courses. He was unable to find any predictors that correlated significantly with his second criterion, persistence in the MBA program (Lewis, 1964).

In a 1965 study, Mittman and Lewis investigated the relationships between the other five predictors used in Lewis' 1964 study and the criteria of GSBT Verbal and Quantitative scores. Stepwise regression revealed that the only background variable to correlate with the verbal test score was the number of undergraduate semester hours taken in business courses. This correlation coefficient was .65 at the .05 significance level. The relatively strong relationship between verbal scores on the GSBT and the number of undergraduate business courses, demonstrates that the GSBT is a good achievement test. Undergraduate department and undergraduate major were also found to be

significantly correlated with the quantitative test (Mittman and Lewis, 1965).

Criterion problems

The hardest part of a predictive validation study is to obtain suitable criterion data (Chronbach, 1970). Aptitude tests are normally validated against grade-point averages, but these are not always a stable criterion because of differences in raters and in evaluation criteria.

> If a criterion measure is not stable, not consistent, it will be impossible for any test or other predictor to relate well with it. (Womer, 1968)
>
> Another problem with the criterion is the possibility that it is biased in favor of certain groups of people. If the criterion is affected by factors unrelated to the attribute it is designed to measure, it may be biased. (Womer, 1968)

Travers and Wallace advised that graduate schools monitor the stability of average grades from year to year and from department to department. They found that in one engineering school, useful prediction of GGPA was impossible because of its large variability (1950).

Departmental differences

Madaus and Walsh (1965) investigated the differences between departments. They found that department sizes were strongly related to the sizes of the correlation coefficients achieved for those departments. They observed higher correlations in larger departments than they saw in smaller ones, or in the university as a whole. When GRE Verbal and Quantitative scores were used to predict GGPA, correlations ranged from .22 for the entire sample of 569 students to .69 for a single department (N = 68). Based on their findings the researchers wrote:

> It would appear, therefore, that the size of N is a definite factor relative to whether or not a significant relationship is found between the dependent and independent variables. The findings of this study lead one to the conclusion that the practice of grouping departments for predictive purposes should not be employed. No matter how logical the grouping appears to be, the results are likely to be of limited utility. (Madaus and Walsh, 1965)

Grouping departments to increase sample size, based on judgement appears to be counter-productive. Jensen recognized that differences between departments occured because student abilities and grading practices vary between departments. If the groups statistically differ with respect to the relevant variables, variability tends to increase and correlation coefficients tend to decrease (Jensen, 1953). On the other hand, if the differences

between groups were tested using statistical methods, and groups of similar programs formed, the correlation coefficents should not decrease significantly.

In their review of 189 GRE validity studies, Livingston and Turner observed that within the group of 41 departments having less than 25 students, variations in the correlation coefficients were noticeably large. This occurred between departments and within the same department from year to year. Their analysis of these variations caused the authors to state:

> Individual departments differ widely in the correlations of GRE scores with FYA, but these differences may mainly be the result of small sample instability. (1982)

FYA in the previous quote refers to first year graduate grade-point average. It seems likely that differences in departmental grading criteria may have been partly responsible for the inter-departmental differences, and that the effect of small sample instability is better shown in the year to year intra-departmental fluctuations.

Lin and Humphreys used a sample selection strategy to reduce the effects of differential grading standards. They selected three particular graduate departments because:

> they attract somewhat similar students, have large numbers of graduate students, and have faculties that have more or less maintained reliable and valid standards of graduate and undergraduate grading. This last criteria barred a large number of departments from consideration. (1977, p.250)

They found that the academic performance of students with high test scores and UGPA's was more stable than those with lower test scores and grades, and the performance of better students was more predictable (1977, p. 252).

The selection ratio

Prediction of academic performance is an important topic for research. While the consequences of inaccurate decisions or policies are serious, accurate prediction of success in graduate school is elusive. The variability of undergraduate and graduate grades, the effects of restriction in range and small sample sizes have consistently been cited as factors contributing to prediction problems. Researchers have investigated a number of predictors and criteria with mixed results. In general, background variables have had, at best, moderate correlations with GGPA. The most commonly chosen predictors, UGPA, GRE test scores, and GMAT test scores have usually demonstrated statistically significant relationships with GGPA, but have seldom yielded correlations researchers consider neccesary. In this

respect many promising research efforts have been abandoned too quickly. As Brogden (1946) demonstrated, even a small improvement in selection can be valuable in many situations.

A critical element in determining the value of criterion-related validity research has gone unmentioned by many researchers. This element is the selection ratio. The selection ratio can be computed by dividing the number of selected applicants by the total number of applicants. Taylor and Russell (1939) demonstrated convincingly that the usefulness of tests with a validity of less than .70 increases more and more as the selection ratio becomes smaller.

They developed a series of tables that depict the relationships between the selection ratio, the proportion of individuals rated satisfactory (before the use of a predictor or prediction model), and the validity of the predictor or prediction model. By using the appropriate table, a researcher can estimate the benefits that can be derived from the use of a predictor or prediction model, based on a validity estimate that reflects the influence of the selection ratio. For example, if 50% of the present students in a graduate school were successful, a selection ratio of .5 was used, and the validity of a new test was .6, the tables show that 94% of those selected would be

successful. The substantial increase in the proportion of successful students from 50%, which would be expected if all applicants were admitted, to 94%, if a test with a validity of .6 was used and the selection ratio remained constant, shows the powerful effects of the selection ratio (Taylor and Russell, 1939, pp.570-578).

<u>Summary</u>

A wide range of approaches has been used in criterion-related validity studies. A lack of agreement concerning the relevant variables, and appropriate techniques for analyzing their inter-relationships has resulted in a large number of exploratory investigations and few in-depth studies. Even when research has identified promising techniques, or potentially important variables, later researchers have seldom attempted to incorporate them in their own studies. The results paint a clear picture of what has not worked in a variety of specific situations, but leave only a vague impression of what may be useful in general application.

It is clear that there is room for improvement in the prediction of graduate school success. It is equally apparent that reliance on published data to support the use of a particular test or prediction model cannot be justified. There are important differences between schools

and between departments within them, that make local validity research necessary. Nearly every researcher has agreed in one respect: continuing research and empirical studies of criterion-related validity are needed. Based on the material reported here, it is clear that these recommendations, at least, are valid.

Research hypotheses

1. The correlations of the predictor variables with GGPA vary between AFIT master's degree programs. In at least some cases the differences between program correlation coefficients are statistically significant.

2. The correlations computed for the entire sample are lower than at least some of those computed for individual programs.

3. When groups are formed based on statistically similar predictor/criterion relationships, and multi-variate regression models are developed for those groups, the prediction models developed for the groups contain different sets of predictors and different predictor weights.

4. Graduate Record Examinations test scores, Graduate Management Admissions Test scores, and undergraduate grade-point average are valid predictors of graduate grade-point average.

5. Background variables such as commissioned years of service (CYRS), enlisted years of service (EYRS), and number of undergraduate math courses (NMAT) add to the accuracy of one or more of the prediction models.

6. The models developed in this study are more accurate than AFIT's current selection procedures.

CHAPTER II

METHODS

Subjects

The subjects in this study include all resident AFIT master's degree students in the School of Systems and Logistics and the School of Engineering who attended AFIT between 1977 and 1982, inclusive. The information collected includes relevant predictor, criterion, and biographical data for approximately 98% of the total population group. The total data base includes 2170 cases. Demographic information is contained in Appendix A.

Variable definitions

For convenience, abbreviated variable names will be used throughout the remainder of this thesis. The variable names are defined below.

GMTT	GMAT composite score
GMTV	GMAT Verbal subtest score
GMTQ	GMAT Quantitative subtest score
GRET	The sum of the GRE verbal and quantitative subtests
GREV	GRE Verbal subtest score
GREQ	GRE Quantitative subtest score
GREA	GRE Analytical subtest score
EYRS	Enlisted years of service
CYRS	Commissioned years of service
NMAT	Number of undergraduate mathematics courses
TOEF	Test of English as a Foreign Language score
UGPA	Undergraduate grade-point average
GGPA	Graduate grade-point average

Data analysis

In the first step of the analysis, correlation matrices containing all of the variables were calculated for the entire sample and for each of the 17 AFIT resident master's degree programs using the Statistical Package for the Social Sciences (SPSS) Pearson Corr program (Nie, et al, 1975). The matrices were calculated using pair-wise deletion of missing values so that each correlation coefficient would be based on the largest possible sample size. This was necessary because the number of cases with missing values was very large. For example, only 1330 of the 2170 cases (61.3%) contained GRE data.

The data base contained information on non-graduates, late-graduates, and graduates, however it did not contain information on applicants who had not been selected for AFIT resident master's degree programs. Because the mean scores of the selected group and the non-selected group differ for those variables used in making selection decisions, it is necessary to consider the effects of restriction in range. Restriction in range attenuates the correlation coefficients between the predictors and the criterion. In cases where only a small proportion of applicants are selected, the attenuation can be significant (Thorndike, 1949, pp.169-176). In the groups studied here, there is a direct restriction on the predictor variables as a result of the selection process.

The initial correlation coefficients were corrected to take this attenuation into account using the formula derived by Thorndike (1949, p.173).

Frequently, problems associated with small sample instability have been mentioned as limiting factors in criterion-related validity research. The range of correlation coefficients for the master's programs studied here was large, indicating that the same problems may be present. To reduce the effects of small sample instability, an effort was made to determine whether or not some of the programs could be combined to form larger, but still homogeneous, groups. A preliminary inspection of the correlation matrices showed that only a few of the predictor variables consistently correlated with GGPA at a .05 significance level in more than half of the 17 programs. The matrices were examined to determine which of the variables were significantly related to the criterion in the largest number of programs with the following results:

Predictor Variable Name	Number of Programs Significant
GRET	13
GREQ	13
UGPA	10
GREV	6
EYRS	6
GREA	5
CYRS	5
GMTQ	3
NMAT	3
GMTT	2
GMTV	2
TOEF	0

It was decided to use the subset of predictors containing the GRET, GREQ, GREV, and UGPA as the basis for comparing the predictor/criterion relationships between programs due to missing data among the other predictors. Statistically significant predictor/criterion relationships were compared across programs using the method outlined in Cohen and Cohen (1975, pp.50-52). Because the sampling distribution of non-zero correlation coefficients is skewed, it was necessary to use Fischer's Z Transformation to convert the distribution of independent correlation coefficients to a nearly normal distribution. The transformed values were tested using a procedure very similar to a T-test (Cohen and Cohen, 1975).

The observed significance levels (p-values) calculated in this process were tabulated in matrix form. The table of p-values can be found in Appendix B. Although

the number of possible program combinations was large, the requirement that the programs be similar with respect to at least two of their predictor/criterion correlations eliminated a great number of possible combinations. In the end, five homogeneous program groups were formed. In these groups the predictor/criterion relationships for two or more predictors were not significantly different ($p < .05$). Correlation matrices for these five program groups, and fo the entire sample, can be found in Appendix C. The resulting program groups are reported in Chapter III.

<u>Developing prediction models</u>

Stepwise multiple regression was used to calculate prediction models for each of the five groups. This method has the advantage of weighting each predictor in direct proportion to its correlation with the criterion and in inverse proportion to its correlation with other predictors. The highest weight is assigned to the predictor with the highest validity and the least overlap with other predictors in the model. Since optimum weights are developed for each predictor, the multiple correlation coefficient that results has the highest validity that is possible for that set of predictors (Anastasi, 1976, pp.180-183).

Since some of the independent variables used in the regression models were highly intercorrelated, the likelihood of multi-colinearity inducing a blocking effect on the introduction of subsequent independent variables into the model had to be considered. To prevent a variable that was highly correlated with both the dependent variable and the other independent variables from reducing the overall multiple correlation coefficient, the independent variables were systematically dropped from the equation. This procedure has been suggested both as a means to identify a multi-colinearity problem if one exists, and to eliminate its effects on the calculation of a regression equation (Nie, et al, 1975, pp.340-341).

The "best" model for each of the groups was chosen based on a comparison of multiple correlation coefficients. These models are reported in Chapter III.

CHAPTER III

RESULTS

This chapter contains four sections. In section one, evidence supporting the validity of the predictor variables is presented. Section two contains a brief analysis of the validity of the procedure currently used in selecting AFIT students, and reports the outcome of this procedure. In the third section the prediction models developed in this research project are listed, and their usefulness is discussed. The fourth section is a short economic analysis of the benefits that could result from the use of the prediction models developed in this study.

Validity of the predictors

The correlations between each of the twelve predictor variables and GGPA are shown in Table 1. These correlation coefficients were computed for the entire sample (N = 2170), but because data on some of the variables were missing from many of the cases, the individual correlations are based on smaller sample sizes. For some of the variables the reduction in sample size is very large. A full correlation matrix can be found in Appendix C.

Table 1

Correlations of predictors with GGPA (entire sample)

VARIABLE:	GMTT	GMTV	GMTQ	GRET
CORRELATION:	.440	.465	.285	.315
SAMPLE SIZE:	386	381	381	1330
SIGNIFICANCE:	0.00	0.00	0.00	0.00
VARIABLE:	**GREV**	**GREQ**	**GREA**	**EYRS**
CORRELATION:	.163	.351	.401	-.31
SAMPLE SIZE:	1330	1330	456	342
SIGNIFICANCE:	0.00	0.00	0.00	0.00
VARIABLE:	**CYRS**	**NMAT**	**TOEF**	**UGPA**
CORRELATION:	.191	-.05	.402	.187
SAMPLE SIZE:	1976	2090	28	2168
SIGNIFICANCE:	0.00	0.03	0.06	0.00

Table 1 shows that the GMAT tests and the GRE Analytical test <u>are</u> correlated with GGPA when all AFIT master's degree programs are grouped together. In addition, it shows that their correlations with GGPA are stronger than those of GRET, GREV, and UGPA which are being used by the AFIT Registrar's office as the primary indicators for the engineering master's programs, and as alternates for the logistics school programs.

Comparing the correlations

The correlations reported in Table 1 were based on a sample containing 17 different master's degree programs. It is logical to assume that they represent a middle ground between the highest and lowest correlations found in individual programs. A comparison of the correlation coefficients that were calculated for the 17 master's degree programs supports this hypothesis. Substantial differences in the relationships between the predictor variables and GGPA were observed, even when programs that appear to be somewhat similar on the surface were compared. For example correlation coefficients for GREV with GGPA ranged from -.447 (N = 115) in the Aeronautical Engineering program to .674 (N = 36) in the Systems Engineering program.

Some of the differences in correlations can be attributed to the instability of correlation coefficients in small samples, although most of the sample sizes reported here for <u>individual programs</u> are equal to or larger than those commonly reported in the literature. Sample correlation coefficients for each of the predictor/criterion relationships were compared with the object of combining programs into statistically similar groups. As a result of this process 15 of the 17 programs were combined into 5 groups. Members of each of these

groups had correlation coefficients for two or more predictor/criterion relationships that were not significantly different.

This process demonstrated that some programs could be grouped together to reduce the effects of small sample instability without significantly degrading predictor/criterion relationships that were observed in the individual programs, and added support to the hypothesis that statistical combination of groups would reveal similarities not intuitively obvious.

<u>Predictor/criterion correlations for program groups</u>

. Tables 2 through 6 show the correlations between the relevant predictors and GGPA for each of these groups. These correlations demonstrate the validity of the predictor/criterion relationships in the program groups. With the exceptions of GRET, GREV, GREQ, and UGPA which are reported in every case for purpose of comparison, predictors that did not correlate with GGPA at the .10 significance level are not included in the tables.

Table 2

Correlations of predictors with GGPA (Group #1)

ASTRONAUTICAL ENGINEERING
SYSTEMS MANAGEMENT
SYSTEMS ENGINEERING

VARIABLE:	GRET	GREV	GREQ	UGPA
CORRELATION:	.658	.538	.622	-.05
SAMPLE SIZE:	167	167	167	296
SIGNIFICANCE:	0.00	0.00	0.00	0.27

Table 3

Correlations of predictors with GGPA (Group #2)

STRATEGY AND TACTICS (O.R.)
ELECTRICAL ENGINEERING OPTICS
ELECTRICAL ENGINEERING

VARIABLE:	GRET	GREV	GREQ
CORRELATION:	.308	.129	.367
SAMPLE SIZE:	285	285	285
SIGNIFICANCE:	0.00	0.06	0.00

VARIABLE:	GREA	CYRS	UGPA
CORRELATION:	.163	.167	.341
SAMPLE SIZE:	117	422	429
SIGNIFICANCE:	0.10	0.00	0.00

Table 4

Correlations of predictors with GGPA (Group #3)

LOGISTICS MANAGEMENT
ENGINEERING MANAGEMENT
CONTRACTING MANAGEMENT
ACQUISITION MANAGEMENT

VARIABLE:	GRET	GREV	GREQ	GREA
CORRELATION:	.372	.233	.324	.531
SAMPLE SIZE:	515	515	515	166
SIGNIFICANCE:	0.00	0.00	0.00	0.00

VARIABLE:	CYRS	NMAT	UGPA
CORRELATION:	.139	.160	.158
SAMPLE SIZE:	457	484	470
SIGNIFICANCE:	0.01	0.01	0.14

Table 5

Correlations of predictors with GGPA (Group #4)

AERONAUTICAL ENGINEERING
ENGINEERING PHYSICS
OPERATIONS RESEARCH

VARIABLES:	GRET	GREV	GREQ
CORRELATIONS:	.308	.129	.367
SAMPLE SIZE:	285	285	285
SIGNIFICANCE:	0.00	0.06	0.00

VARIABLES:	GREA	CYRS	UGPA
CORRELATIONS:	.163	.167	.341
SAMPLE SIZE:	117	270	277
SIGNIFICANCE:	0.10	0.02	0.00

Table 6

Correlations of predictors with GGPA (Group #5)

COMPUTER SCIENCE
NUCLEAR EFFECTS ENGINEERING

VARIABLES:	GRET	GREV	GREQ	UGPA
CORRELATION:	.322	.010	.492	.273
SAMPLE SIZE:	142	142	142	181
SIGNIFICANCE:	0.00	0.00	0.00	0.00

Description of present admissions procedures.

The Air Force uses a three-step process in screening potential students for programs under AFIT's jurisdiction. In the first step, academic records are reviewed by AFIT's evaluators and the names of all academically eligible officers are transmitted to the Air Force Military Personnel Center (MPC). AFIT's academic evaluation is a continuous process. Since AFIT is the repository for all active duty Air Force officer educational records, these records are forwarded to AFIT shortly after an officer is commissioned. When AFIT receives them, the records are screened to determine whether or not the officer meets the eligibility criteria for admission to the AFIT programs that are related to his/her career field or past academic experience.

Those officers whose academic records are above average will normally be classified as eligible for AFIT programs as a result of the initial evaluation. In this manner, officers who have not formally applied for admission are "centrally identified." Officers may also become eligible for AFIT programs by requesting evaluation (volunteering). AFIT's position is that volunteers are better motivated to succeed in AFIT graduate programs.

These individuals need not have above average academic records, but they must meet AFIT's minimum criteria. AFIT provides educational counseling to volunteers who do not meet eligibility criteria. If additional transcripts showing that deficiencies have been corrected are forwarded to AFIT at a later date, the officer's records are re-evaluated and eligibility may be granted at that time. The names of all officers qualified by either of these processes are placed on an AFIT eligibility listing. Updated versions of this computer listing are transmitted to MPC periodically (Bigelow, 1983).

The current listing shows that approximately 13,000 officers have attained eligibility status through the processes described above (Air Force Institute of Technology, 1983). This can be contrasted with the number of Air Force officers who have not yet earned a masters degree. According to the Air Force Magazine (May, 1983), there are 51,190 line officers in this category. Of that total, only 25.3% are included in the group that AFIT considers eligible.

Although the minimum eligibility criteria vary from program to program, in general they consist of the following:

1. Undergraduate GPA of 2.5 or higher
2. GRE verbal and quantitative test scores

totalling 1000. GMAT scores of 500 or better are prefered for some programs, but GRE scores are acceptable.

3. A minimum number of math courses (depending on degree type).

4. Grades of "C" or better in required courses. (U.S. Air Force Manual 50-5, Volume I, para 4-15, 4-16, 1981)

In the second step of the process, career field managers at MPC review the military records of eligible officers under their purview to determine which of them are available for an assignment, have the required job experience, and have acceptable performance rati..gs. Once this review is completed, selection folders containing the relevant portions of the academic and military records of the officers eligible for AFIT are prepared for review by MPC's selection board. Since each of the career field managers acts independently, and has a different quota to fill, it is doubtful that this part of the screening process is conducted uniformly. Minimum criteria is specified by Air Force Manual 50-5, Volume I, para 4-15 (a):

> a.Milita · Availability. Officers must:
> (1) _e medically unrestricted for worldwide duty.
> (2) Be serving in the grade of colonel or below.
> (3) Have a competitive military record.

(4) Be available for reassignment.

(5) Have at least 3 years intervening service since last PCS education on the date of class entry. (United States Air Force, 1981)

In addition to those requirements, officers must also meet the following criteria, which are among those specified in AFM 50-5, Volume I, para 15 (c):

c. Assignment availability:

(1) On-station requirements:

(a) Normally, the AFIT entry date provides for a minimum of 24 months on station before school entry.

(b) Officers serving on or projected to serve on overseas tours are scheduled for school entry to coincide with their DEROS. (United States Air Force, 1981)

The final phase of the screening process occurs when the officer's military and educational records are evaluated by a selection board of senior officers. According to AFM 50-5, Volume I:

A continuous selection board convenes begining in July (each year) to consider line of the Air Force applicants and centrally identified officers below the rank of colonel for AFIT entry during the next fiscal year. Out of cycle selections are made throughout the year from late volunteers and PCS available officers to fill any remaining vacancies.(para 4-22 (a), 1981)

The selection process is highly competitive and considers overall academic military performance and post-AFIT assignment suitability. Factors include promotability, career progression, prior academic and assignment experience and the qualifications of the individual to perform in positions requiring the education to be obtained through AFIT. The selection process is designed to select officers whose potential contribution after graduation will most benefit the Air Force. (para 4-22 (b), 1981)

This board functions differently from a military promotion

board, and is closely related to the assignment process. MPC's career field managers, whose primary interest is in the assignment process, have a significant influence on AFIT selection board decisions (Bigelow, 1983).

Determining the procedure's validity

The number of people involved in the screening process makes analyzing the current procedures a difficult task. For the purpose of this thesis, analyzing the result of the process is a better starting point. If success at AFIT is defined in terms of graduation on time, the data collected in this study shows that 90.4% of those selected for AFIT meet that criterion. Of those who did not graduate on time, 26.9% eventually completed their degree requirements. In other words, 92.99% of those who attended AFIT resident master's degree programs between 1977 and 1982 (inclusive) have completed graduation requirements. This is a very respectable figure, when compared to the graduation rates normally found in civilian graduate institutions. However, there are other factors that must be considered.

The selection ratio has a direct bearing on the results of a selection process. During the period of 1977 to 1982 inclusive, an average of 362 students were selected for AFIT resident master's degree programs each year.

Assuming that the number of eligibles has remained fairly constant over that period, a useful estimate of the selection ratio for this time period is 362/13000 or 2.7%. The actual selection ratio must be lower than 2.7% because that estimate includes only officers who are eligible for AFIT (25.3% of the population). A selection ratio of this size significantly enhances the accuracy of a selection process. Given an estimate of the graduation rate that would have occurred had no screening process been used, it is possible to estimate the validity of the selection process itself.

The graduation rate that would have occurred had there been no selection was estimated at 69%. This figure assumes that essential undergraduate prerequisite courses or course sequences had been completed and that the applicant's undergraduate degree is in the required field to qualify him/her for graduate study in an AFIT resident master's program. It does not reflect the use of cut-off scores for GRE or GMAT test or for UGPA. The method used to estimate the graduation rate is shown in Appendix D.

Using Taylor-Russell tables, the validity of the Air Force's selection process was estimated at .35. The fact that this level of validity can produce a 90.4% graduation rate demonstrates the benefits that result from use of a very low selection ratio. Further examination of

the Taylor-Russell tables shows that a selection model with a validity of .65 or better would increase the graduation rate to 99%. Relevant Taylor-Russell tables are contained in Appendix E.

In the first phase of its selection process the Air Force uses multiple-criteria cut-off scores to screen applicants. Selection procedures of this kind may be useful when the number of applicants is large, and the evaluation methods are relatively inexpensive, but they are problematic. The effect of the multiple cut-off scores is to eliminate individuals from consideration based on subjective criteria weighting systems, rather than more objective statistically derived formulae. If any of the eligibility criteria are set too high, or are irrelevant, a significant portion of potentially successful applicants can be excluded.

The large number of missing values in the data indicates that the multiple cut-off score criteria are not being applied uniformly. An applicant who formally requests that his/her eligibility for AFIT programs be evaluated is required to submit GRE or GMAT scores. Other officers, whose initial eligibility was determined based on the other criteria (i.e., those that were centrally selected), may be selected for AFIT without consideration of GRE/GMAT scores. This situation occurs because "the

AFIT selection cycle does not always tie in with the ETS testing cycle", according to Mr. C. P. Bigelow, Chief of AFIT's Evaluation and Counseling Section (1983).

The academic evaluation of those officers who are not volunteers (who have not forwarded test scores to AFIT) is based largely on UGPA. Considering that the correlations found between UGPA and GGPA in this study range from .15 to .34, predictions based on UGPA alone are questionable, especially when better information is available. This practice results in the use of a different set of predictors (and predictor weights) for those who have furnished AFIT with test score data and those who have not. However unavoidable they may be, these circumstances result in a more stringent screening of volunteers for AFIT than of non-volunteers, benefiting the non-volunteers. A procedure that uniformly applied standard cut-off scores for all criteria to all applicants would at least insure that all applicants were considered on the same basis.

You may recall the study by Borg (1963) that was discussed earlier. He found that the use of a single cut-off score for the GRE Verbal test would have denied admission to 41 successful students as well as 21 unsuccessful students. Since this process would have eliminated nearly twice as many successful students as unsuccessful students from consideration, he recommended

against its use. Because of their cumulative effects, the Air Force's use of multiple cut-off scores may be producing even more undesirable effects. Determining the extent of these effects would be extremely difficult because the use of multiple cut-off scores results in relationships that are non-linear. Predictions based on multiple cut-off criteria involve complicated mathematics and can only be made for a given set of scores.

<u>Best prediction models</u>

Prediction models were developed using a step-wise linear regression program. A series of regression models were calculated. To insure that the best combination of variables was used, each of the predictor variables was dropped from the equation in turn. In most cases, at least one variable was dropped from the regression model before the highest multiple R was achieved. The "best models" shown below were chosen on the basis of a comparison of multiple R's.

Table 7

Multiple regression equation
(entire sample using cases with both GRE and GMAT)

PREDICTOR	WEIGHT
GMTT	+0.002798929
GRET	+0.002024500
GREV	-0.002382224
GMTQ	-0.026734060
UGPA	-0.081269790
CONSTANT	+1.999044000
MULTIPLE R =	0.51692
SAMPLE SIZE =	108

Table 8

Multiple regression equation
(entire sample using cases with GMAT)

PREDICTOR	WEIGHT
GMTV	+0.021495910
GMTQ	+0.005746557
UGPA	+0.035720140
CONSTANT	+2.601640000
MULTIPLE R =	0.47809
SAMPLE SIZE =	364

Table 9

Multiple regression equation
(entire sample using cases with GRE)

PREDICTOR	WEIGHT
GREA	+0.001622894
UGPA	+0.139886900
GREV	-0.001924470
GRET	+0.001060860
CONSTANT	+1.941677000
MULTIPLE R =	0.49388
SAMPLE SIZE =	419

Table 10

Multiple regression equation (Group #1)

ASTRONAUTICAL ENGINEERING
SYSTEMS MANAGEMENT
SYSTEMS ENGINEERING

PREDICTOR	WEIGHT
GRET	-0.004151110
UGPA	-0.255726700
GREQ	+0.007465280
GREV	+0.005263321
CONSTANT	+1.364894000
MULTIPLE R =	0.71036
SAMPLE SIZE =	161

TABLE 11

Multiple regression equation (Group #2)

STRATEGY AND TACTICS (O.R.)
ELECTRICAL ENG (OPTICS)
ELECTRICAL ENGINEERING

PREDICTOR	WEIGHT
GREQ	+0.001219023
UGPA	+0.421528200
CYRS	+0.045127900
GREA	+0.001149079
GREV	-0.001095729
CONSTANT	+0.981167700

MULTIPLE R =	0.5762
SAMPLE SIZE =	117

Table 12

Multiple regression equation (Group #3)

LOGISTICS MANAGEMENT ENGINEERING MANAGEMENT CONTRACTING MANAGEMENT ACQUISITION MANAGEMENT	
PREDICTOR	WEIGHT
GMTV	+0.008191018
GMTQ	+0.008754303
UGPA	+0.225509900
CYRS	+0.026706010
NMAT	+0.036887990
CONSTANT	+2.122747000
MULTIPLE R =	0.55204
SAMPLE SIZE =	187

Table 13

Multiple regression equation (Group #4)

AERONAUTICAL ENGINEERING ENGINEERING PHYSICS OPERATIONS RESEARCH	
PREDICTOR	WEIGHT
UGPA	+0.477362800
GREQ	-0.005689860
GREV	+0.010004080
GRET	+0.008815480
CONSTANT	+0.404797700
MULTIPLE R =	0.69005
SAMPLE SIZE =	245

Table 14

Multiple regression equation (Group #5)

COMPUTER SCIENCE NUCLEAR EFFECTS	
PREDICTOR	WEIGHT
GRET	+0.003032275
GREV	-0.004153440
UGPA	+0.175798500
CONSTANT	+1.495748000
MULTIPLE R =	0.64412
SAMPLE SIZE =	133

Group #3 contains all the programs in the School of Systems and Logistics, with the exception of Systems Management. It was the only group in which there were enough cases with GMAT scores to permit a comparison of models based on GRE and GMAT. The model based on cases with GMAT scores was the better of the two. (In the GRE based model, multiple R = .49, N = 127.)

GREA is a relatively new subtest of the GRE. Because it is new, this variable was missing from many cases. Inclusion of it in a model reduced the sample size. However, GREA's correlation with GGPA was generally one of the highest for each of the groups. For this reason, and the need to establish its contribution to prediction in the various programs, it was included in as many prediction models as possible. In every case, it increased the multiple correlation coefficients over those found without it.

These prediction models confirm the findings reported in the first section of this chapter. That is, they show that each of the program groups has its own unique "best" set of predictors and predictor weights. Furthermore, the different weights the variables take on in the linear models demonstrate the importance of using statistical means to establish a selection formula.

Economic analysis

In Chapter I the costs associated with sponsoring a student in AFIT resident master's degree programs were mentioned as justification for this research. For convenience, the figures are repeated here.

Engineering school cost = \$82,892.68 per student

Logistics school cost = \$67,258.66 per student

In this study it was determined that 145 Engineering School students and 63 Logistics School students failed to graduate with their classmates. If you assume that each non-graduate (or late graduate) represents a total loss on the Air Force's investment, the cost of selection errors can be determined easily.

145 x \$82,892.68	=	\$12,019,439.00
63 x \$67,258.66	=	\$ 4,237,295.60
Total	=	\$16,256,734.60

The assumption that a non-graduate represents a total loss on the investment, seems more reasonable when you consider that a student <u>could have been selected</u> who would have graduated.

In the previous section which examined the validity of current selection procedures, it was noted that a prediction model with a validity of .65 or better would

increase the graduation rate from 90.4% to 99%. The validities of the models developed for the five program groups in this study range from .55 to .71. The uniform application of these models should increase the graduation rate to between 97% and 100%. When you consider that the Air Force's loss through incorrect admissions decisions averaged more than $2.7 million per year over the six years included in this study, investing a fraction of that amount to implement a new selection strategy makes good sense.

CHAPTER IV

DISCUSSION AND CONCLUSIONS

A review of the hypotheses

The first hypothesis stated that statistically significant differences in predictor/criterion correlations would be found when the correlations were compared across AFIT programs. The correlations between the predictor variables and GGPA varied significantly between AFIT master's degree programs, adding support to the findings of Madaus and Walsh (1965). Furthermore, for many AFIT master's degree programs, the predictor/GGPA correlations were not significant at the .05 or even .1 significance level. This finding was unexpected, and indicates that the use of variables that appear to be logically related to the criterion is unsound. Until the validity of a predictor is demonstrated statistically it should not be used.

The second hypothesis is related to the first. It stated that correlation coefficients for the entire sample would be lower than some of those computed for individual programs. It was also supported.

The third hypothesis, that the regression models developed for statistically combined program groups would differ in terms of their predictors and predictor weights was supported.

The fourth hypothesis, that GRE scores, GMAT scores, and UGPA are valid predictors of GGPA can only be supported for some of the programs studied. For example, the correlation of GREV with GGPA is statistically significant at the .05 level in only 6 of the 17 master's programs. Even when GREV is a statistically significant predictor of GGPA the correlations differ widely from program to program. The correlations for GREV with GGPA range from -.447 in the Aeronautical Engineering program to .674 in the Systems Engineering program, but the correlation for the entire sample was .163. Assuming that the correlation is the same for all three groups is a serious error. The other predictors followed the same pattern as GREV, though not to the same extreme. These predictors should be used only for specific situations in which their correlations with the criterion are known.

The fifth hypothesis, that background variables would add to the accuracy of at least one of the prediction models was supported. CYRS entered two of the final prediction models and NMAT entered one.

The last hypothesis, that the models developed in this study are more accurate than AFIT's current selection procedure was also supported. This finding was expected. The literature contains a great deal of support for the use of statistical procedures in solving problems of this kind, and it offers very little support for the use of judgement or intuition.

Discussion

This study demonstrates the concept of differential validity. The correlation coefficients calculated show that the differences between programs in a single graduate school can be significant. The prediction models developed through multiple regression add additional support to that finding, and show that different sets of predictors are appropriate for different programs. It is evident from the range of correlations calculated for the various programs that success in some programs can be predicted more accurately than others.

The use of statistical procedures to compare the relationships between predictors and GGPA within the 17 programs showed that the differences between some groups for as many as three predictor/criterion relationships were not statistically significant. The benefits of grouping programs in this manner, rather than through clinical

inference, are demonstrated by the relatively high multiple correlation coefficients that were achieved for the grouped programs. This technique holds promise for many situations in which large individual samples do not exist. As far as can be determined, this is the first validity study to combine groups statistically for prediction of success in graduate school.

Other findings

Some interesting variables were examined. These include commissioned years (CYRS), and enlisted years (EYRS). CYRS provided low but statistically significant correlations in 5 of the academic programs. EYRS was statistically significant only in the 6 of the 17 academic programs where the proportion of officers with prior enlisted service time was fairly large. In 5 of these programs its correlations with GGPA ranged from -.51 to -.75, indicating that officers with enlisted experience may be at a substantial disadvantage in graduate school. Where the numbers are great enough for it to assume significance, this variable could be very useful.

The effects of moderator variables were investigated early in this study. This line of research was dropped because the key predictors had a large number of missing values and selecting cases based on moderator

variables drastically reduced sample sizes. However, some interesting effects were noted. The relationships between the predictors and GGPA were stronger for service academy graduates than for those who obtained undergraduate degrees from other sources. The performance of Second Lieutenants in engineering programs was well below average performance in those programs. Predictor/criterion correlations for married officers were higher than for single officers in most of the programs. While moderator variables were not especially useful in this study, these findings indicate that there may be a great many variables that are useful in predicting performance.

Conclusions

AFIT's present selection accuracy is better than what could be expected at a private university. The validity study described in this thesis relied on well established psychological measurement techniques, but it combined them in a new way. As a result, it has shown that selecting students through these methods could result in even better selection accuracy than presently exists.

Selecting students for graduate school is no simple task. The relationship between success and past performance varies from one situation to another. This study has demonstrated that variability exists between

correlates of success in resident master's degree programs at the Air Force Institute of Technology. It has established the validity of current selection procedures, of five proposed selection models, and of several predictor variables. Since the predictor data are already contained in the academic or military records of potential students, it offers the Air Force some new toc to aid the selection process. More importantly it has shown that a selection procedure that uses multiple cut-off scores only for absolutely essential prerequisites, and uses a linear model incorporating other relevant variables to predict performance in the criterion would result in improved selection.

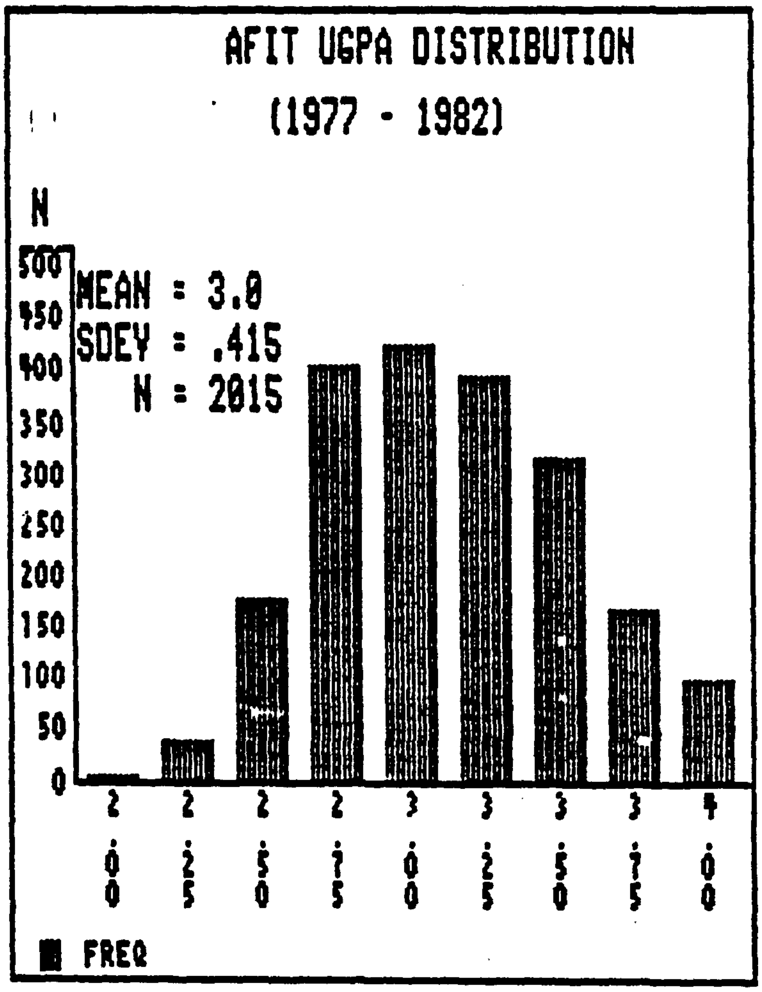
The structure of the Air Force personnel assignment system and the dual procedure for determining eligibility for AFIT programs complicate the selection process. The concept of selecting those best qualified for graduate education is certainly appropriate, but it may be difficult in this environment. Because the selection system is a sub-set of the assignment system, some compromises are probably necessary. Early notification of eligibles, including those centrally identified, and the requirement that <u>all</u> these officers submit test scores before receiving an assignment to an AFIT graduate program would improve the process.

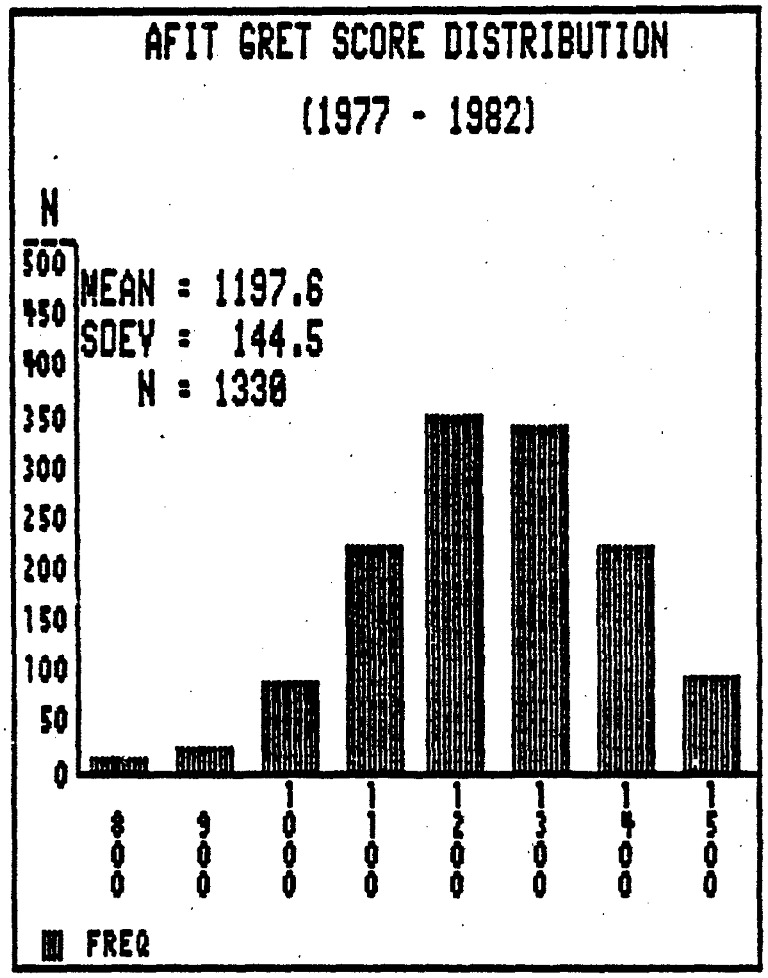
The School of Systems and Logistics has recently begun emphasizing the requirement that all applicants submit test scores. This effort primarily influences the 1984 class and subsequent classes. It is a step in the right direction. Models such as those developed in this study are effective and relatively easy to use if all the data are available. If data are not available the practical benefits they offer are limited.

This study points to a larger issue. That issue is the human cost involved in selecting some applicants and rejecting others. The cost of choosing someone who will eventually fail is high for that individual. By the same token, the cost of rejecting someone who could have suceeded is large. In many cases limited resources, differences in ability, and external constraints make this cost unavoidable, but it should be minimized whenever possible. It may be difficult to translate into dollars and cents, but it is real.

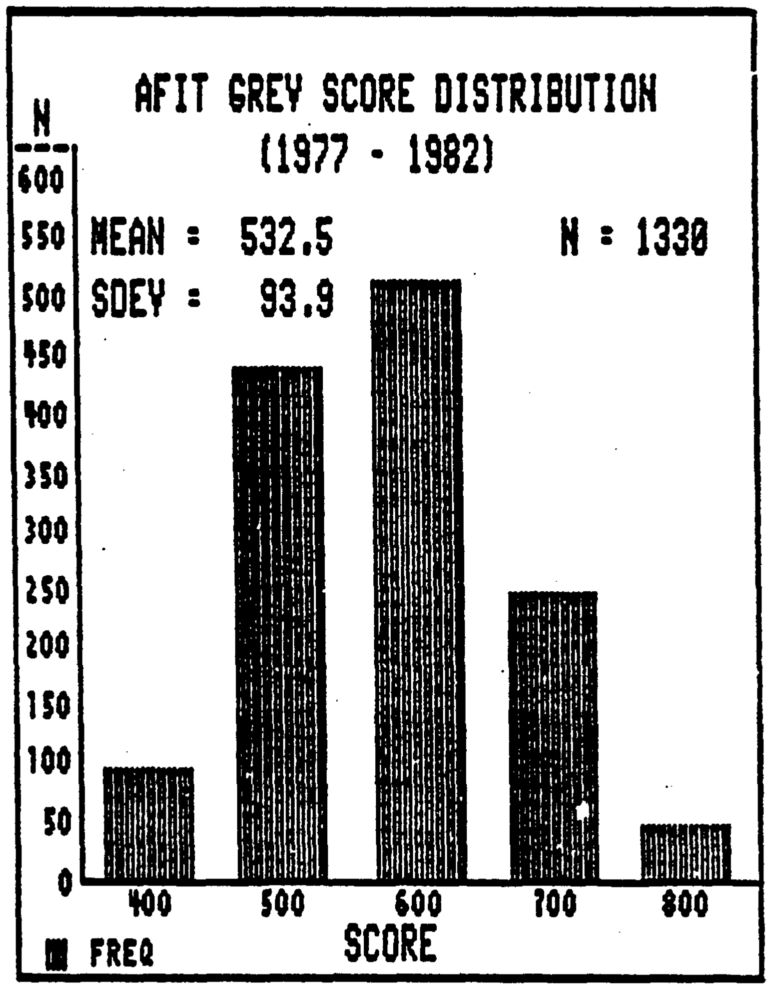
APPENDIX A

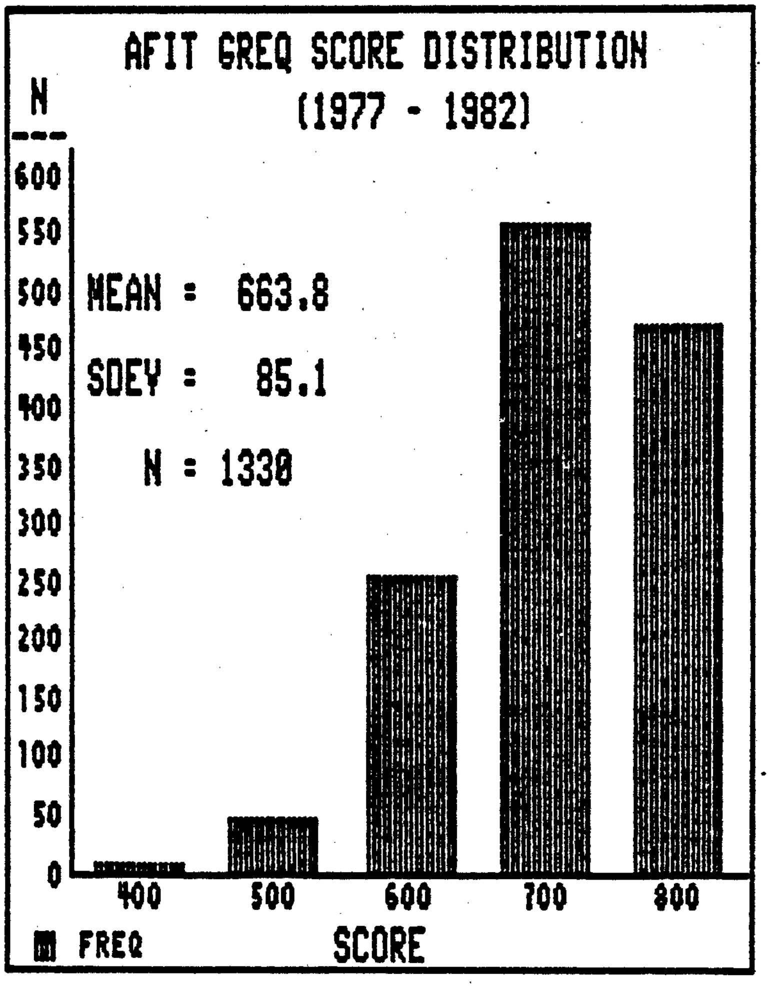
DEMOGRAPHIC INFORMATION

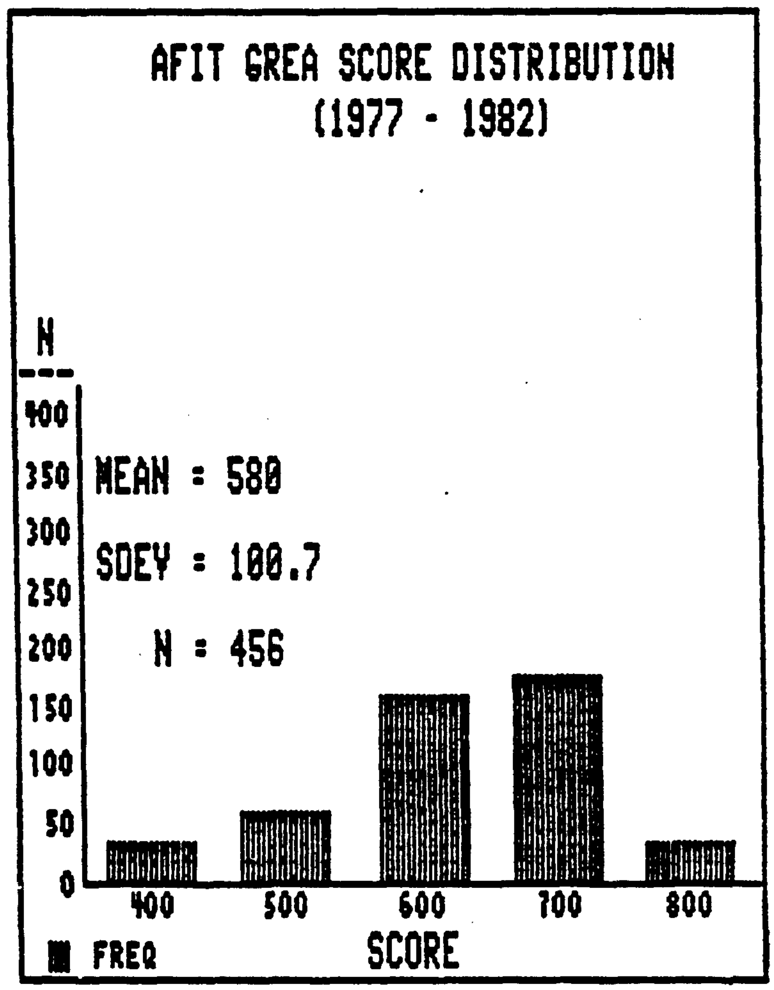

AFIT UGPA DISTRIBUTION
(1977 - 1982)
N
MEAN = 3.0
SDEV = .415
N = 2015
500
450
400
350
300
250
200
150
100
50
0
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
FREQ

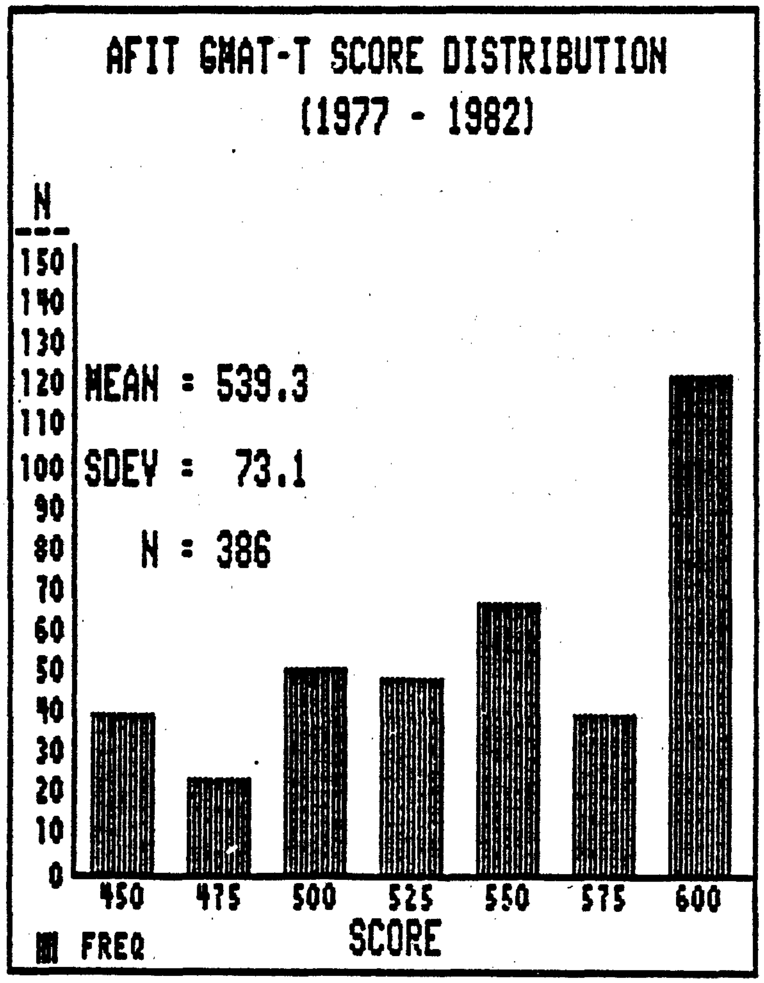


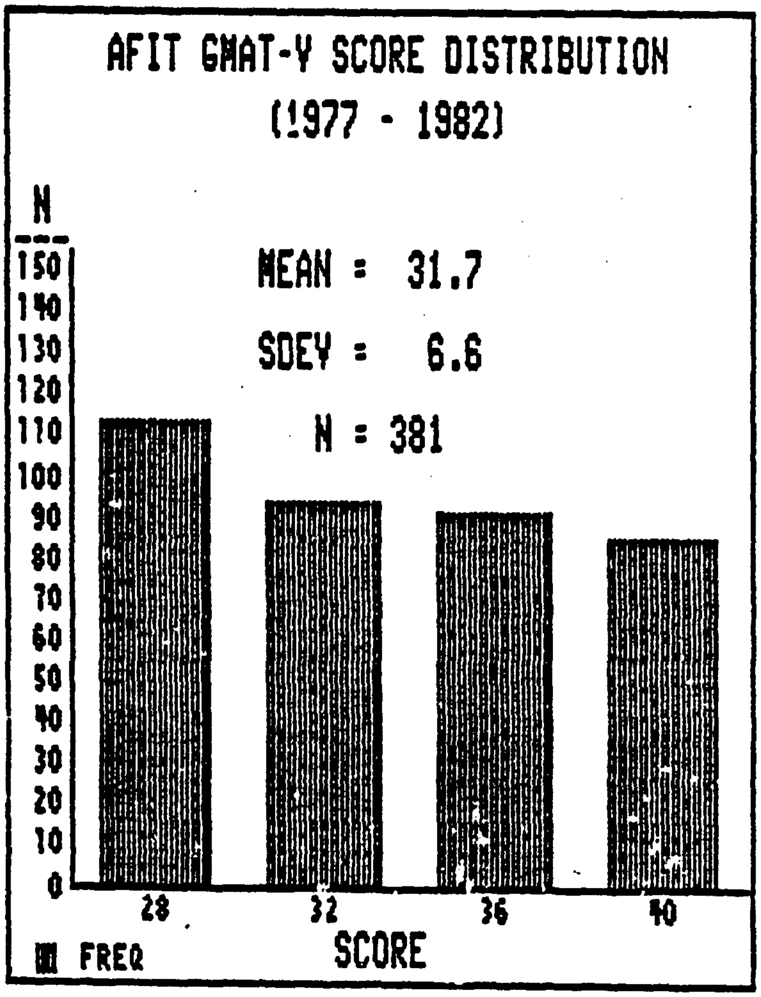
AFIT GRET SCORE DISTRIBUTION
(1977 - 1982)
N
500
450
400
350
300
250
200
150
100
50
0
MEAN = 1197.6
SDEV = 144.5
N = 1330
800
900
1000
1100
1200
1300
1400
1500
FREQ

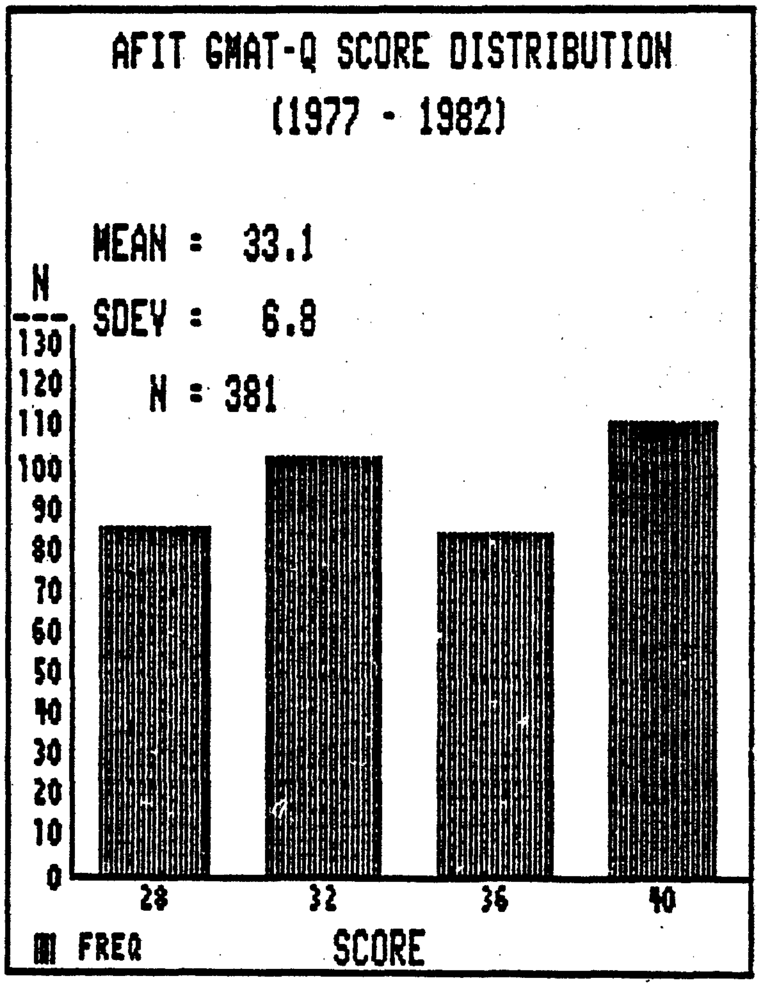

AFIT GREY SCORE DISTRIBUTION
(1977 - 1982)
N
600
550
500
450
400
350
300
250
200
150
100
50
0
MEAN = 532.5
SDEV = 93.9
N = 1330
400
500
600
700
800
SCORE
FREQ

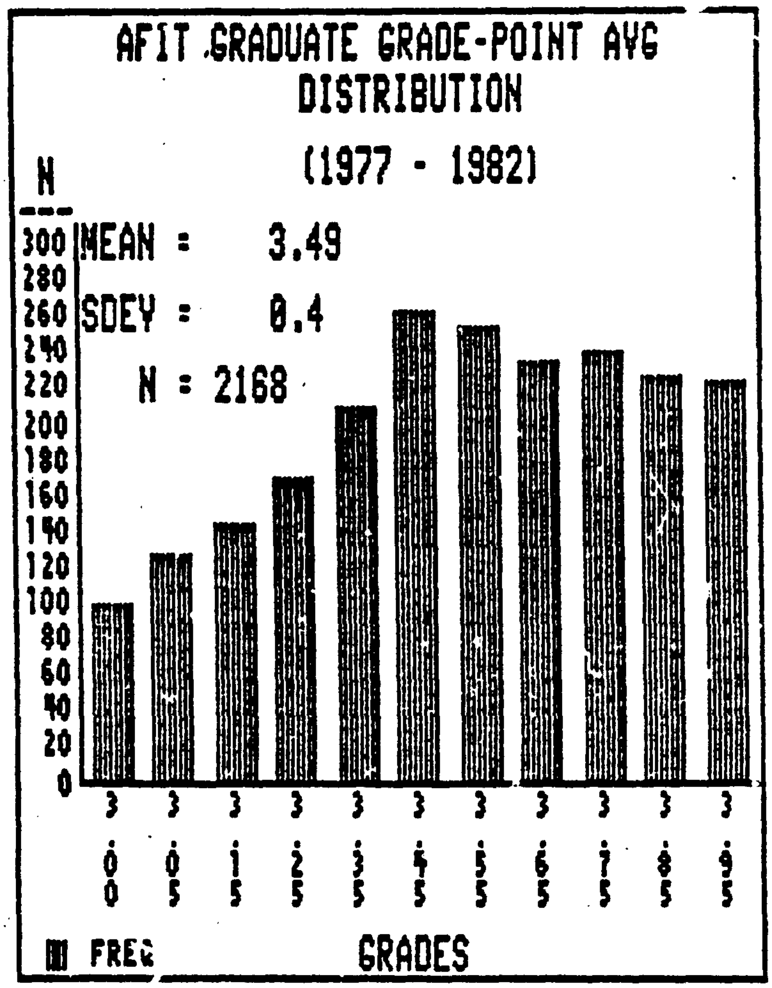

AFIT GREQ SCORE DISTRIBUTION
(1977 - 1982)
N
600
550
500
450
400
350
300
250
200
150
100
50
0
MEAN = 663.8
SDEV = 85.1
N = 1330
400
500
600
700
800
FREQ
SCORE


AFIT GREA SCORE DISTRIBUTION
(1977 - 1982)
N
400
350
300
250
200
150
100
50
0
MEAN = 580
SDEV = 100.7
N = 456
400
500
600
700
800
FREQ
SCORE

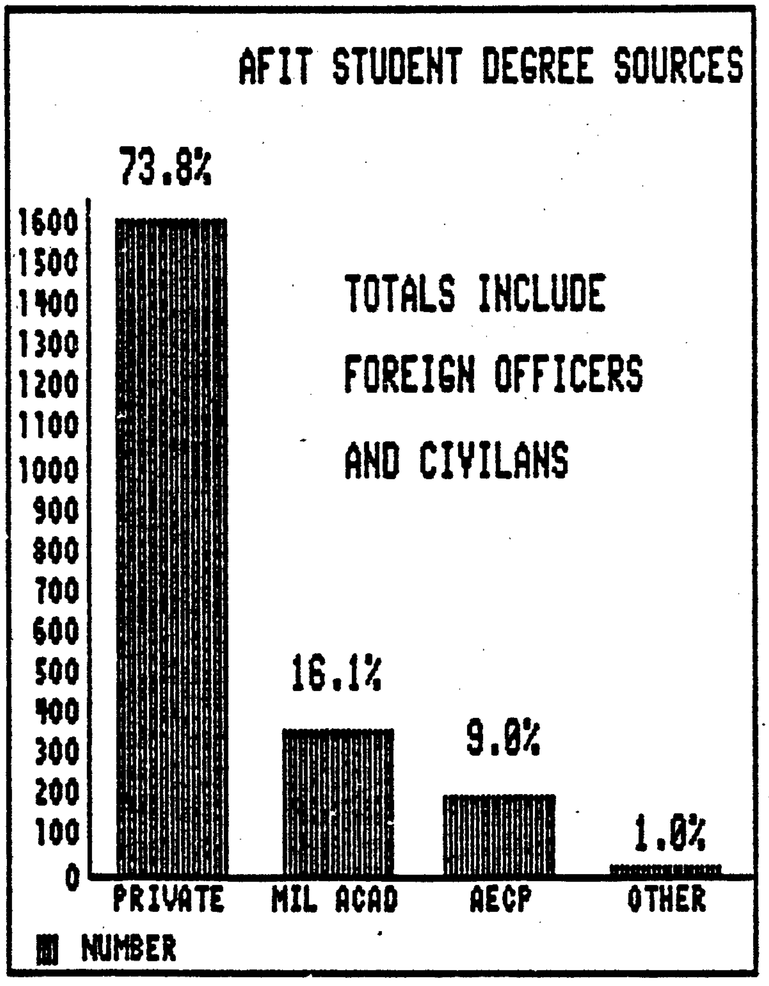

AFIT GMAT-T SCORE DISTRIBUTION
(1977 - 1982)
N
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
MEAN = 539.3
SDEV = 73.1
N = 386
450
475
500
525
550
575
600
FREQ
SCORE

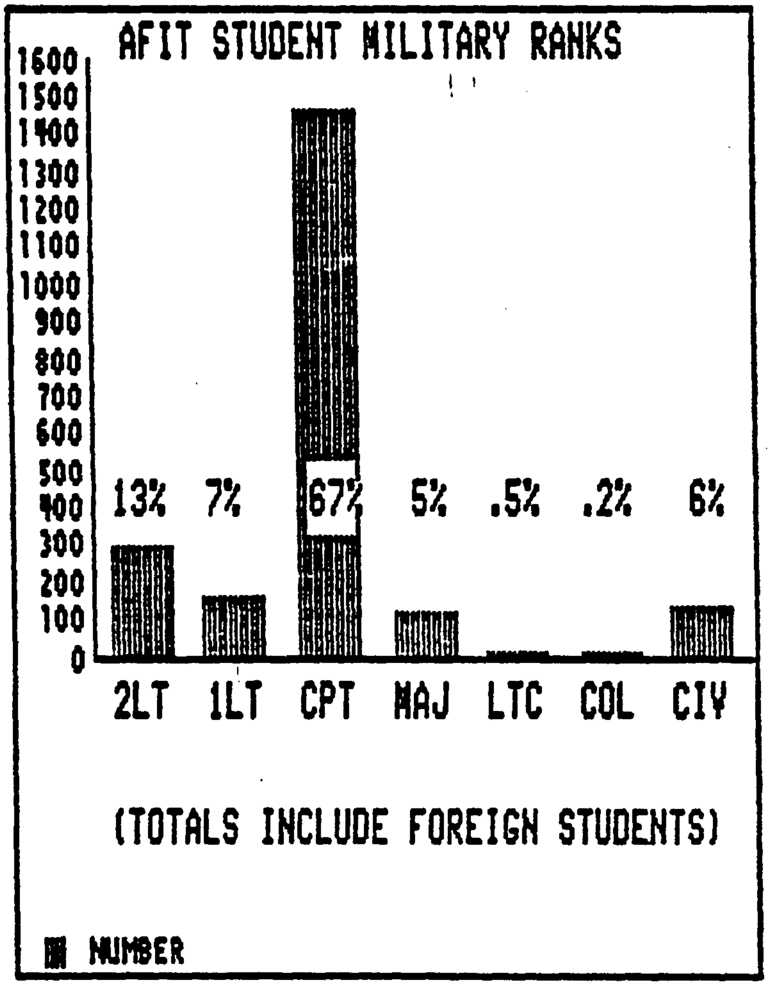

AFIT GMAT-V SCORE DISTRIBUTION
(1977 - 1982)
N
MEAN = 31.7
SDEV = 6.6
N = 381
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
28
32
36
40
SCORE
FREQ


AFIT GMAT-Q SCORE DISTRIBUTION
(1977 - 1982)
MEAN = 33.1
SDEV = 6.8
N = 381
N
130
120
110
100
90
80
70
60
50
40
30
20
10
0
28
32
36
40
FREQ
SCORE



AFIT GRADUATE GRADE-POINT AVG
DISTRIBUTION
(1977 - 1982)
N
MEAN = 3.49
SDEV = 0.4
N = 2168
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
3.00
3.05
3.15
3.25
3.35
3.45
3.55
3.65
3.75
3.85
3.95
FREQ
GRADES


AFIT STUDENT DEGREE SOURCES
73.8%
16.1%
9.0%
1.0%
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
TOTALS INCLUDE
FOREIGN OFFICERS
AND CIVILANS
PRIVATE
MIL ACAD
AECP
OTHER
NUMBER


AFIT STUDENT MILITARY RANKS
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
13%
7%
67%
5%
.5%
.2%
6%
2LT
1LT
CPT
MAJ
LTC
COL
CIV
(TOTALS INCLUDE FOREIGN STUDENTS)
NUMBER

APPENDIX B

CORRELATION TEST MATRICES

TABLE OF P VALUES

CALCULATED IN GRET/GGPA CORRELATION TEST

BETWEEN PROGRAM CORRELATION COEFFICIENTS

FOR GRET WITH GGPA

	Comp	EEng	GEO	GEP	Nucl	OpsR
Astro	.146	.012	.246	.029	.225	.020
Comp		.250	1.000	.339	.865	.267
EEng			.422	.966	.596	.842
GEO				.459	.885	.401
GEP					.609	.820
Nucl						.540

	SysE	Strat	Contr	EMgt	SysMgt	Log
Astro	.414	.260	.059	.010	.942	.031
Comp	.036	.826	.458	.180	.147	.513
EEng	.003	.271	.472	.689	.009	.500
GEO	.239	.860	.647	.312	.258	.658
GEP	.007	.323	.926	.775	.028	.595
Nucl	.073	.754	.689	.459	.234	.834
OpsR	.005	.270	.834	.881	.018	.454
SysE		.093	.015	.000	.365	.007
Strat			.472	.197	.320	.470
Contr				.723	.059	.734
EMgt					.009	.336
Sys Mgt						.028

If P < Alpha, reject the hypothesis that the two programs are from the same population of students

TABLE OF P VALUES

CALCULATED IN GREV/GGPA CORRELATION TEST

BETWEEN PROGRAM CORRELATION COEFFICIENTS

FOR GREV WITH GGPA

	Aero	Nucl	SysEng	SysMgt	LOG
Astro	.000	.888	.153	.165	.357
Aero		.000	.000	.000	.000
Nucl			.263	.317	.395
SysEng				.764	.014
SysMgt					.003

If P < Alpha, reject the hypothesis that the two programs are from the same population of students

TABLE OF P VALUES

CALCULATED IN GREQ/GGPA CORRELATION TEST

BETWEEN PROGRAM CORRELATION COEFFICIENTS

FOR GREQ WITH GGPA

	Aero	Comp	EEng	GEO	GEP	OpsR
Astro	.035	.289	.002	.101	.007	.004
Aero		.213	.368	.984	.392	.28
Comp			.023	.373	.051	.028
EEng				.524	.88	.696
GEO					.502	.401
GEP						.851

	SysE	Strat	Contr	EMgt	Log	SysMgt
Astro	.656	.212	.051	.003	.004	.303
Aero	.217	.575	.833	.262	.510	.308
Comp	.756	.681	.236	.021	.035	.944
EEng	.061	.207	.660	.702	.729	.060
GEO	.317	.664	.853	.395	.646	.441
GEP	.075	.617	.617	.861	.682	.370
OpsR	.051	.163	.496	.976	.509	.057
SysE		.513	.211	.047	.084	.681
Strat			.513	.153	.276	.749
Contr				.495	.808	.303
EMgt					.496	.048
Log						.068

If P < Alpha, reject the hypothesis that the two programs are from the same population of students

TABLE OF P VALUES

CALCULATED IN UGPA/GGPA CORRELATION TEST

BETWEEN PROGRAM CORRELATION COEFFICIENTS

FOR UGPA WITH GGPA

	Aero	Comp	EEng	GEO	GEP	Nucl
Astro	.972	.822	.741	.410	.210	.789
Aero		.749	.631	.318	.124	.741
Comp			.920	.478	.230	.920
EEng				.478	.204	.960
GEO					.719	.660
GEP						.441

	OpsR	Strat	LOG
Astro	.481	.456	.744
Aero	.389	.366	.698
Comp	.582	.535	.457
EEng	.589	.542	.265
GEO	.849	.936	.166
GEP	.555	.654	.038
OpsR		.916	.196
Strat			.198

If P < Alpha, reject the hypothesis that the two programs are from the same population of students

APPENDIX C

CORRELATION MATRICES

MATRIX OF CORRELATION COEFFICIENTS

ENTIRE AFIT SAMPLE

	GMTT	GMTV	GMTQ	GRET
GMTT	1.000	.918	.863	.803
GMTV		1.000	.543	.655
GMTQ			1.000	.768
GRET				1.000

	GREV	GREQ	UGPA	GGPA
GMTT	.688	.662	.298	.434
GMTV	.664	.381	.238	.400
GMTQ	.501	.792	.273	.366
GRET	.870	.843	.316	.396
GREV	1.000	.425	.128	.262
GREQ		1.000	.432	.364
UGPA			1.000	.145
GGPA				1.000

MATRIX OF CORRELATION COEFFICIENTS

GROUP #1

ASTRONAUTICAL ENGINEERING
SYSTEMS ENGINEERING
SYSTEMS MANAGEMENT

	GRET	GREV	GREQ	UGPA	GGPA
GRET	1.000	.859	.871	.215	.658
GREV		1.000	.505	.087	.538
GREQ			1.000	.257	.622
UGPA				1.000	-.052
GGPA					1.000

MATRIX OF CORRELATION COEFFICIENTS

GROUP #2

STRATEGY AND TACTICS (O.R.)
ELECTRICAL ENGINEERING OPTICS
ELECTRICAL ENGINEERING

	GRET	GREV	GREQ	GREA	CYRS	UGPA	GGPA
GRET	1.000	.878	.863	.636	.067	.212	.308
GREV		1.000	.501	.661	.046	.140	.129
GREQ			1.000	.472	.037	.216	.367
GREA				1.000	-.208	.033	.163
CYRS					1.000	-.374	.167
UGPA						1.000	.341
GGPA							1.000

MATRIX OF CORRELATION COEFFICIENTS

GROUP #3

LOGISTICS MANAGEMENT
ENGINEERING MANAGEMENT
CONTRACTING MANAGEMENT
ACQUISITION MANAGEMENT

	GRET	GREV	GREQ	GREA
GRET	1.000	.827	.847	.693
GREV		1.000	.333	.636
GREQ			1.000	.465
GREA				1.000

	CYRS	NMAT	UGPA	GGPA
GRET	.018	.335	.102	.372
GREV	.024	.080	.158	.23[illegible]
GREQ	.012	.454	-.011	.374
GREA	.107	.108	.162	.531
CYRS	1.000	.044	-.514	.139
NMAT		1.000	-.240	.160
UGPA			1.000	.158
GGPA				1.000

MATRIX OF CORRELATION COEFFICIENTS

GROUP #4
AERONAUTICAL ENGINEERING
ENGINEERING PHYSICS
OPERATIONS RESEARCH

	GRET	GREV	GREQ	UGPA	GGPA
GRET	1.000	.889	.700	.274	.130
GREV		1.000	.313	.238	-.071
GREQ			1.000	.216	.312
UGPA				1.000	.493
GGPA					1.000

MATRIX OF CORRELATION COEFFICIENTS

GROUP #5
COMPUTER SCIENCE
NUCLEAR ENGINEERING

	GRET	GREV	GREQ	UGPA	GGPA
GRET	1.000	.856	.858	.230	.322
GREV		1.000	.439	.159	.001
GREQ			1.000	.215	.492
UGPA				1.000	.273
GGPA					1.000

APPENDIX D

METHOD USED TO ESTIMATE THE VALIDITY OF CURRENT AFIT SELECTION PROCEDURES

METHOD USED TO ESTIMATE THE VALIDITY OF CURRENT AFIT SELECTION PROCEDURES

Mean GREV and GREQ scores for the unrestricted group were obtained from an Educational Testing Service report furnished to AFIT (Educational Testing Service, 1981). The means were calculated using all scores reported to AFIT between October, 1980 and October, 1981. They were based on data from non-selectees as well as selectees. The ratio of the scores from this unrestricted group to those of the students selected for AFIT (the restricted group) provided an index that was used to estimate what the mean GGPA would have been had all applicants that met essential criteria been accepted.

The mean (unrestricted) GGPA was estimated by multiplying the AFIT group GGPA by both of these indexes, summing the products, and dividing by 2. This method was used to insure that the estimate would be conservative.

This figure was converted to a Z score by subtracting the critical GGPA (3.0) and dividing by the unrestricted standard deviation, which was calculated in the same manner. The Z score was then converted into a corresponding area of the normal curve. This area of the normal curve (.19) was added to the area on the other side

of the normal curve (0.5). The result is the estimate of the percentage of (unrestricted) students that would have earned a GGPA of 3.0 or better (69%).

With this information, and the selection ratio, the Taylor-Russell tables in Appendix E can be used to estimate the validity of AFIT's current selection procedures. The table for .70 shows that with a selection ratio of .05, the validity of the current procedures must fall between .30 and .35.

COMPUTATIONS

Step 1

$$\text{Ratio \#1} = \frac{\text{Unrestricted Group GREV Mean Score}}{\text{AFIT Student Group GREV Mean Score}}$$

$$= \frac{520}{532.5} = .976$$

$$\text{Ratio \#2} = \frac{\text{Unrestricted Group GREQ Mean Score}}{\text{AFIT Student Group GREQ Mean Score}}$$

$$= \frac{609}{663.8} = .917$$

Step 2

(Ratio #1) X (AFIT Mean GGPA) =

(.976) X (3.4793) = 3.3957

(Ratio #2) X (AFIT Mean GGPA) =

(.917) X (3.4793) = 3.1905

Step 3

3.3957 + 3.1905 = 6.5862

(6.5862) X (0.5) = 3.293

3.293 = Estimated Mean GGPA for an Unrestricted Group of Students

Step 4

$$\frac{\text{Estimated Mean GGPA - Pass Fail Score}}{\text{Unrestricted GGPA Standard Deviation}} = \text{Z score}$$

$$\frac{3.293 - 3.0}{.5895} = .4970\ (Z)$$

.4970 (Z) = area of the normal curve = .19

.5 + .19 = .69

.69 = the percentage of unrestricted applicants who could be expected to pass given that multiple cut-off scores were not used except to establish that absolutely essential prerequisites had been satsified

APPENDIX E

TAYLOR-RUSSELL TABLES

Proportion of Employees Considered Satisfactory = .70
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.70	.70	70	.70	.70	.70	.70	.70	.70	.70	.70
.05	.73	.73		.72	.72	.71	.71	.71	.71	.70	.70
.10	.77	.76		.74	.73	.73	.72	.72	.71	.71	.70
.15	.80	.79	.77	.76	.75	.74	.73	.73	.72	.71	.71
.20	.83	.81	.79	.78	.77	.76	.75	.74	.73	.71	.71
.25	.86	.84	.81	.80	.78	.77	.76	.75	.73	.72	.71
.30	.88	.86	.84	.82	.80	.78	.77	.75	.74	.72	.71
.35	.91	.89	.86	.83	.82	.80	.78	.76	.75	.73	.71
.40	.93	.91	.88	.85	.83	.81	.79	.77	.75	.73	.72
.45	.94	.93	.90	.87	.85	.83	.81	.78	.76	.73	.72
.50	.96	.94	.91	.89	.87	.84	.82	.80	.77	.74	.72
.55	.97	.96	.93	.91	.88	.86	.83	.81	.78	.74	.72
.60	.98	.97	.95	.92	.90	.87	.85	.82	.79	.75	.73
.65	.99	.98	.96	.94	.92	.89	.86	.83	.80	.75	.73
.70	1.00	.99	.97	.96	.93	.91	.88	.84	.80	.76	.73
.75	1.00	1.00	.98	.97	.95	.92	.89	.86	.81	.76	.73
.80	1.00	1.00	.99	.98	.97	.94	.91	.87	.82	.77	.73
.85	1.00	1.00	1.00	.99	.98	.96	.93	.89	.84	.77	.74
.90	1.00	1.00	1.00	1.00	.99	.98	.95	.91	.85	.78	.74
.95	1.00	1.00	1.00	1.00	1.00	.99	.98	.94	.86	.78	.74
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.88	.78	.74

Proportion of Employees Considered Satisfactory = .80
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.80	.80	.80	.80	.80	.80	.80	.80	.80	.80	.80
.05	.83	.82	.82	.82	.81	.81	.81	.81	.81	.80	.80
.10	.85	.85	.84	.83	.83	.82	.82	.81	.81	.81	.80
.15	.88	.87	.86	.85	.84	.83	.83	.82	.82	.81	.81
.20	.90	.89	.87	.86	.85	.84	.84	.83	.82	.81	.81
.25	.92	.91	.89	.88	.87	.86	.85	.84	.83	.82	.81
.30	.94	.92	.90	.89	.88	.87	.86	.84	.83	.82	.81
.35	.95	.94	.92	.90	.89	.89	.87	.85	.84	.82	.81
.40	.96	.95	.93	.92	.90	.89	.88	.86	.85	.83	.82
.45	.97	.96	.95	.93	.92	.90	.89	.87	.85	.83	.82
.50	.98	.97	.96	.94	.93	.91	.90	.88	.86	.84	.82
.55	.99	.98	.97	.95	.94	.92	.91	.89	.87	.84	.82
.60	.99	.99	.98	.96	.95	.94	.92	.90	.87	.84	.83
.65	1.00	.99	.98	.97	.96	.95	.93	.91	.88	.85	.83
.70	1.00	1.00	.99	.98	.97	.96	.94	.92	.89	.85	.83
.75	1.00	1.00	1.00	.99	.98	.97	.95	.93	.90	.86	.83
.80	1.00	1.00	1.00	1.00	.99	.98	.96	.94	.91	.87	.84
.85	1.00	1.00	1.00	1.00	1.00	.99	.98	.96	.92	.87	.84
.90	1.00	1.00	1.00	1.00	1.00	1.00	.99	.97	.94	.88	.84
.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	.96	.89	.84
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.89	.84

[Source: Taylor and Russell, 1939, p.576]

SELECTED BIBLIOGRAPHY

A. REFERENCES CITED

Air Force Institute of Technology. AFIT Education Newsletter, Air Force Institute of Technology/RR, AURP 53-2, Wright-Patterson AFB OH, 1982.

Air Force Institute of Technology. AFIT Electees/Tentative Selectees and EWI Eligibles (WYA3), Unpublished computer listing, Air Force Institute of Technology/RR, Wright-Patterson AFB OH, 10 July 1983.

Air Force Institute of Technology. Financial Report 096CR, Unpublished funds report, Air Force Institute of Technology/ACB, Wright-Patterson AFB OH, 1981.

Air Force Magazine, An Air Force almanac: the United States Air Force in facts and figures, Washington DC: *The Air Force Association* , May 1983, 169.

American Psychological Association, *Standards for educational and psychological tests and manuals* , Washington DC: American Psychological Association, 1966.

Anastasi, A., *Psychological Testing* , New York: Mac Millan, 1976.

Baird, L. L., Comparative prediction of first-year graduate and professional school grades in six fields, *Educational and Psychological Measurement* , 1975, 35, 941-946.

Bigelow, C. P., Chief, Evaluation and Counseling Section, AFIT/RR, Wright-Patterson AFB OH. Personal Interview. 16 August 1983.

Borg, W. R., GRE aptitude scores as predictors of GPA for graduate students in education, *Educational and Psychological Measurement* , 1963, 23, (2), 379-382.

Breaugh, J. A. & Mann, R. B., The utility of discriminant analysis for predicting graduation from a master of business administration program, Educational and Psychological Measurement , 1981, 41, 495-501.

Brogden, H. E., On the interpretation of the correlation coefficient as a measure of predictive efficiency, The Journal of Educational Psychology , 1946, 37, (2), 65-76.

Camp, J. & Clawson, T., The relationship between the graduate record examinations aptitude test and grade point average in a master of arts in counseling program, Educational and Psychological Measurement , 1979, 39, 429-431.

Chronbach, L. J., Essentials of Psychological Testing , 4th ed., New York: Harper and Row Publishers, 1970.

Cohen, J. & Cohen, P., Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences , New York: Halsted Press, 1975.

Covert R. W. & Chansky, N. M., The moderator effect of undergraduate grade point average on the prediction of success in graduate education, Educational and Psychological Measurement , 1975, 35, 947-950.

Cureton, E. E., Validity, reliability, and baloney, Educational and Psychological Measurement , 1950, 10, 94-96.

Educational Testing Service, GRE 1981-1982: guide to the use of the graduate record examinations , Princeton NJ: Educational Testing Service, 1981.

Educational Testing Service, Graduate institution summary statistics report, Unpublished computer listing, Princeton NJ: Educational Testing Service, 1981.

Furst, E. J., Theoretical problems in the selection of students for professional schools, Educational and Psychological Measurement , 1950, (10), 945-952.

Furst, E. J. & Roelfs, P., Validation of the graduate record examinations and the miller's analogies test in a doctoral program in education, Educational and Psychological Measurement , 1979, (39), 145-151.

Green, B. F., A primer of testing, American Psychologist , 1981, 36, (10), 1001-1011.

Graduate Management Admissions Council, Guide to the use of GMAT scores 1982-1983 , Princeton N. J.: Educational Testing Service, 1982.

Guion, R. M., Personnel Testing , New York: McGraw-Hill, Inc., 1965.

Hecht, L. W. & Powers, D. E., The predictive validity of preadmission measures in graduate management education: three years of the GMAC validity study service , Princeton, N.J.: Educational Testing Service, 1982.

Jensen, R. E., Predicting scholastic achievement of first-year graduate students, Educational and Psychological Measurement , 1953, (13), 323-329.

Knapp, J. & Hamilton, I. B., The effect of non-standard undergraduate assessment and reporting practices on the graduate admissions process , Princeton, N.J.: Educational Testing Service, 1978.

Lewis, J. W., The relationship of selected variables to achievement and persistence in a masters program in business education, Educational and Psychological Measurement , 1960, 20, (4), 847-851.

Lin P. & Humphreys, L. G., Predictions of academic performance in graduate and professional school, Applied Psychological Measurement , 1977, 1, (2), 249-257.

Livingston, S. A. & Turner, N. J., Effectiveness of the graduate record examinations for predicting first year grades: 1980-1981 summary report of the graduate record examinations validity study service , Princeton, N.J.: Educational Testing Service, 1982.

Madaus, G. F. & Walsh, J. J., Departmental differentials in the predictive validity of the graduate record examination aptitude tests, Educational and Psychological Measurement , 1965, 25, (4), 1105-1110.

Marston, A. R., It is time to reconsider the graduate record examination, American Psychologist, 1971, 26, 653-655.

Mehrabian, A., Undergraduate ability factors in relationship to graduate performance, Educational and Psychological Measurement, 1969, 29, 409-419.

Mittman, A. & Lewis, J. W., Correlates of achievement on the admissions test for graduate study in business, Educational and Psychological Measurement, 1965, 25, (2), 585-588.

Nagi, J. L., Predictive validity of the graduate record examination and the miller analogies test, Educational and Psychological Measurement, 1975, 35, 471-472.

Nie, N.H., Hull C. H., Jenkins J. G., Steinbrenner, K., & Brent, D. H., SPSS: Statistical Package for the Social Sciences, 2nd ed., New York: McGraw-Hill Book Company, 1975.

Robertson, M. & Nielsen, W., The graduate record examination and selection of graduate students, American Psychologist, 1961, (10), 648-650.

Sawyer, J. T., Measurement and prediction, clinical and statistical, Psychological Bulletin, 1966, 66, (3), 178-200.

Taylor, H. C. & Russell, J. T., The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables, Journal of Applied Psychology, 1939, 23, 565-578.

Thacker, A. J. & Williams, R. E., The relationship of the graduate record examination to grade point average and success in graduate school, Educational and Psychological Measurement, 1974, 34, 939-944.

Thorndike, R. L., Personnel Selection: Test Measurement and Techniques, New York: John Wiley and Sons, Inc., 1949.

Travers, R. M., Personnel selection and classification as a laboratory science, Educational and Psychological Measurement, 1956, 16, 195-208.

Travers, R. M. & Wallace, W. L., The assessment of the academic aptitude of the graduate student, _Educational and Psychological Measurement_ , 1950, 10, 371-379.

Traxler, A. E., Tests for graduate students, _Journal of Higher Education_ , 1952, 23, 473-482.

U.S. Department of the Air Force. USAF Formal Schools Catalog, AFM 50-5, Volume 1, Washington DC: Government Printing Office, 1981.

Willingham, W. W., _Predicting success in graduate education_ , in papers presented to the graduate record examination board research service at the 12th annual meeting of the council of graduate schools, Princeton NJ: Educational Testing Service, 1981.

Womer, F. B., _Basic Concepts in Testing_ , New York: Houghton Mifflin Company, 1968.

B. RELATED SOURCES

Albright, L. E., Glennon, J. R., & Smith, W. J., The uses of psychological tests in industry , Cleveland: Howard Allen, 1963.

Ayers, J. B., Predicting quality point averages in master's degree programs in education, Educational and Psychological Measurement , 1971, 31, 491-495.

Kirnan, J. P. & Geisinger, K. F., The prediction of graduate school success in psychology, Educational and Psychological Measurement , 1981, 41, 815-820.

Lanholm, G. V., Review of the studies employing GRE scores in predicting success in graduate study 1952-1967 , Princeton NJ: Educational Testing Service, 1972.

Saunders, D. R., Moderator variables in prediction, Educational and Psychological Measurement , 1956, 16, 209-222.

Schmidt, F. L., Hunter, J. E., & Urry, V. W., Statistical power in criterion related validity studies, Journal of Applied Psychology , 1976, 61, (4), 473-485.